Almo Jaluta

Probabilistic temperature estimation for a photovoltaic inverter

School of Electrical Engineering

Thesis submitted for examination for the degree of Master of Science in Technology.

Espoo 31.7.2018

Thesis supervisor:

Prof. Marko Hinkkanen

Thesis advisors:

D.Sc. (Tech.) Mikko Qvintus

D.Sc. (Tech.) Lauri Viitasaari

Aalto University
School of Electrical
Engineering

AALTO-YLIOPISTO
SÄHKÖTEKNIIKAN KORKEAKOULU

DIPLOMITYÖN
TIIVISTELMÄ

Tekijä: Almo Jaluta

Työn nimi: Aurinkosähkövaihtosuuntaajan tilastollinen lämpötilan estimointi

Päivämäärä: 31.7.2018 Kieli: Englanti Sivumäärä: 9+59

Sähkötekniikan korkeakoulu

Professuuri: Electric Power and Energy Engineering Koodi: ELEC3024

Valvoja: Prof. Marko Hinkkanen

Ohjaajat: TkT Mikko Qvintus, TkT Lauri Viitasaari

Tämän diplomityön tarkoituksena on tutkia mahdollisuutta tunnistaa vaihtosuuntaajan vioittunut lämpötila-anturi lämpötila-aikasarjoista, muodostaa tilastollinen malli yhdelle lämpötilamittaukselle sekä arvioida, voidaanko vaihtosuuntaajan toimintaa jatkaa tilastollisen mallin avulla lämpötila-anturin vioittuessa.

Materiaalina käytettiin suuriin aurinkovoimaloihin suunnitellun 2 MW:n keskusinvertterin erilaisista kokeista kerättyjä lämpötilamittauksia. Työn tavoitteiden pohjalta muodostettiin tilastollinen menetelmä, joka tunnistaa vioittuneen lämpötila-anturin, simuloi lyhytaikaisia lämpötila-aikasarjoja sekä ennustaa vaihtosuuntaajan toiminnan jatkamisen kannalta, ylittyykö ennalta-asetettu lämpötilaraja. Esitetty malli on rakennettu vioittuneen lämpötila-anturin tunnistavasta lohkosta ja lämpötilaa estimoivasta lohkosta. Ensimmäinen lohko perustuu pääkomponenttianalyysiin, K:n keskiarvon klusterointimenetelmään ja virhe-ellipsiin. Toinen lohko perustuu Markovin ketjuun. Esitetty malli käyttää lähtötietona vain aikaisempia lämpötila-aikasarjoja.

Menetelmän toimivuutta tutkittiin ensin tunnistamalla viallinen lämpötila-anturi sekä vertaamalla estimoitujen lämpötila-aikasarjojen jakaumia historiallisiin lämpötilatietoihin erilaisissa vioittumistapauksissa. Lisäksi menetelmän kykyä ennakoida ennalta-asetetun lämpötilarajan ylittämistä tutkittiin eri esimerkkien avulla. Esitetty menetelmä havaitsi vioittuneet lämpötila-anturit poikkeuksetta. Ennustettujen ja havaittujen lämpötila-aikasarjojen väliset erot olivat hyvin pieniä. Malli pystyi myös ennakoimaan tietyn lämpötilarajan ylittymisen.

Avainsanat: Aurinkosähkö, vaihtosuuntaaja, lämpötila, tilastollinen, estimointi, ennustaminen, pääkomponenttianalyysi, klusterointi, Markov

AALTO UNIVERSITY
SCHOOL OF ELECTRICAL ENGINEERING

ABSTRACT OF THE
MASTER'S THESIS

Author: Almo Jaluta

Title: Probabilistic temperature estimation for a photovoltaic inverter

Date: 31.7.2018 Language: English Number of pages: 9+59

School of Electrical engineering

Professorship: Electric Power and Energy Engineering Code: ELEC3024

Supervisor: Prof. Marko Hinkkanen

Advisors: D.Sc. (Tech.) Mikko Qvintus, D.Sc. (Tech.) Lauri Viitasaari

The purpose of this work is to understand whether a broken temperature sensor can be identified from time series data, if a probabilistic temperature model can be formulated for a single measurement for an outdoor inverter, and whether the inverter can continue converting power under the probabilistic model if the sensor is broken.

Data given for this study were acquired from different experiments during the design and verification of a 2-MW outdoor central inverter for large utility-scale PV power plants. Based on these objectives, probabilistic methodology was constructed to identify outliers in the data, simulate very short-term temperature time series, and evaluate whether a certain temperature threshold is exceeded as a safety measure for continuing inverter operation. The proposed model is constructed of two blocks: an outlier detection block and an estimation block. The first block is based on principal component analysis, K-means and elliptical density estimation. The second block is based on Markov chain. The proposed methodology uses temperature time series data only without knowing the internals of the system.

The proposed model was validated by inputting time-series data containing data from faulty temperature sensors under different failure scenarios, and by comparing simulated temperature time series data to historical temperature data under different cases. Moreover, the simulated time series data were used to verify whether the model can anticipate exceeding a certain temperature threshold. The model always detected the failed sensors. The error metrics of the simulated temperature time series were low. Furthermore, the model anticipated exceeding the given temperature threshold ahead of time.

Keywords: Photovoltaic, inverter, temperature, probabilistic, estimation, forecasting, PCA, clustering, Markov

Preface

First, I wish to thank all the wonderful people in the Department of Mathematics and Systems Analysis whom I have always bothered with my questions for giving me the opportunity to work as a teaching assistant for all these years and for introducing me to the world of probability and statistics. Your enthusiasm and positive energy will continue to inspire me.

I thank my supervisor, Professor Marko Hinkkanen, for his valuable feedback. I would like to thank Mikko Qvintus, one of my instructors, for his insightful comments and for helping me navigate through this project. I also thank Lauri Viitasaari, another of my instructors, for reviewing the mathematical theory and methodology and for his continuous encouragement.

I thank Tomi Riipinen for introducing the research topic of this thesis and ABB Product Group Solar for the opportunity to work on this research topic.

To my family, thank you for your patience throughout the process and for supporting me in all my pursuits in life.

Helsinki, 31.7.2018

Almo Jaluta

Contents

Symbols and abbreviations

Symbols

$\mathbf{A}$	matrix
λ	scalar
$\mathbf{u}$	vector
$\mathbb{R}$	set of all real numbers
μ	population mean
$\bar{x}$	sample mean of variable x
σ_x	standard deviation of random variable X
σ_x^2	variance of random variable X
σ_{xy}	covariance of random variables X and Y
Σ	variance-covariance matrix
$\mathcal{N}(\mu, \sigma^2)$	normal distribution with mean μ and variance σ^2
°C	degrees Celsius
X	random variable
X_{est}	estimated value for random variable X
X_i	$i : th$ observation of random variable X
f_g	grid frequency (Hz)
P	active power (W)
P_{in}	input active power (W)
P_{loss}	active power losses (W)
P_{out}	output active power (W)
P_{ref}	reference active power (W)
Q	reactive power (Var)
Q_{in}	input reactive power (Var)
Q_{out}	output reactive power (Var)
Q_{ref}	reference reactive power (Var)
S	apparent power (VA)
S_{out}	output apparent power (VA)
t	time (s)
T	temperature (°C)
t_{stab}	stabilization time
T_{stab}	stabilization temperature
U_{dc}	dc voltage
U_g	grid voltage (V)
$U_{g,var}$	grid voltage variation (%)

Operators

$\mathrm{E}[X]$	expected value of random variable X
$\mathrm{SD}(X)$	standard deviation of random variable X
$\mathrm{Var}(X)$	variance of random variable X
$\mathrm{Cov}(X, Y)$	covariance of random variables X and Y
$\mathrm{P}(A)$	probability of event A occurring
$\mathrm{P}(A \mid B)$	conditional probability of event A occurring given that event B has occurred
$\mathbf{A}^T$	transpose of a matrix
$\lfloor \rfloor$	floor operator
Δ	difference
$\sum_i^n$	indexed sum
$m \times n$	multiplication
$\in$	set membership
$\ll$	much-less-than sign
$\binom{n}{k}$	binomial coefficient

Abbreviations

2D	two-dimensional
3D	three-dimensional
AC	alternating current
Act	actual
DB	database
DC	direct current
DTMC	discrete time Markov chain
Est	estimate
EUT	equipment under test
IDA	initial data analysis
kW	kilowatt
M	median
MC	Markov chain
MCMC	Markov chain Monte Carlo
MCS	Monte Carlo simulation
MAD	median absolute deviation
MAE	mean absolute error
MVA	multivariate analysis
ML	machine learning
MW	megawatt
PC	principal component
PCA	principal component analysis
PV	photovoltaic
RES	renewable energy source
RMSE	root-mean-square error
SD	standard deviation
SVD	singular value decomposition
TC	thermocouple
TSA	time series analysis
VSI	voltage-source inverter
VST	very short-term
VSTF	very short-term forecasting

1 Introduction

1.1 Background

In photovoltaic (PV) systems, solar energy is harnessed and converted into power for the utility system [1]. PV power generation is one of the fastest growing technologies in the renewable energy sources (RES) domain, and the growth of the technology has been increased by the decreased price of PV panels and support policies in the form of premiums and feed-in tariffs in many countries [2]. However, the fluctuating nature of PV power makes it difficult to achieve its full potential and it is identified as one of the key challenges for massive PV integration [3]. The energy market has difficult-to-meet specifications, including high efficiency, grid code compliance, reliability and long warranty periods [4].

Grid-connected PV systems are the most popular among different PV system topologies, accounting for more than 99% of the total installed PV energy capacity worldwide [4]. In grid-connected PV systems, the power is converted to the grid. Compared to other PV system topologies, the grid-connected systems are more economical and require less maintenance and reinvestment [4]. Figure 1 illustrates the structure of a grid-connected PV system. The system is composed of PV cells, an input filter, a three-phase voltage source inverter (VSI), an output filter and the grid [4]. The PV cells are connected in a series-parallel configuration to generate direct current (DC), and this DC power depends on solar irradiance, temperature and voltage at the solar panels [5]. The generated DC power is then converted into grid-synchronized alternating current (AC) power via a PV inverter.

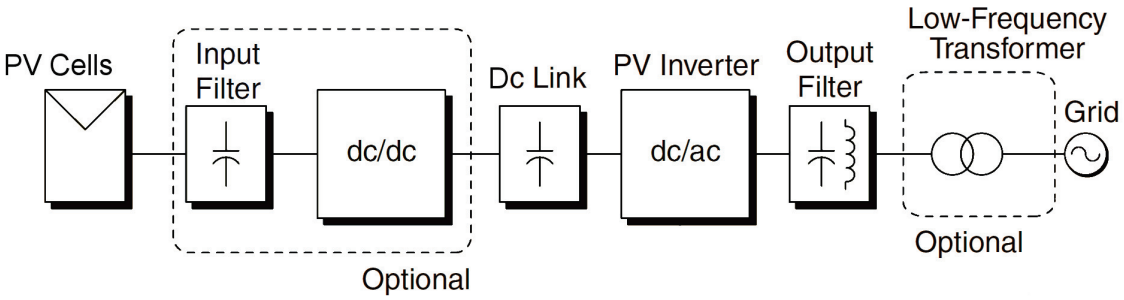


Figure 1: Structure of a grid-connected PV system [4].

The energy production of a PV system depends on the amount of solar irradiation on the panels. However, solar irradiation is not uniform, causing the output power of a PV system to vary substantially over time [6]. This variability in the power production of PV systems poses challenges and uncertainties in grid management and grid stability. When considering the overall reliability of a PV system, the fluctuating nature of the energy produced by the PV system is accompanied by unexpected faults that lead to downtime, preventing PV systems from being reliable sources of power [7]. Adding battery storage to a grid-connected PV system

reduces the fluctuation of the energy produced, however it increases the capital and operation costs significantly [8].

The inverter is the cornerstone of grid-connected PV systems [9]. Therefore, the reliability and availability of the inverter play important roles in grid stability, but they are usually neglected [10]. A recent review of field data provided by PV plant operators indicates that the inverter is the part of the PV system that generates the greatest operation and maintenance costs for PV plant operators [11], and excluding grid effects and lightning, light inverter failures are profound sources for system halts in central inverter topology [10]. The unscheduled downtime leads to lost energy production, and light inverter faults such as software and sensor failures are the predominant causes of power production loss events for inverters [11]. For a PV plant owner, the cost of a light inverter fault, for example, a failed temperature sensor, is equivalent to the cost of service call and spare parts along with the value of the energy that would have been produced during the downtime [12]. Furthermore, industry-wide surveys conducted to determine the requirements and future outlook of power converter reliability across different applications have exposed a need to improve inverter reliability and minimize the unexpected downtime due to light inverter faults [13]. Many sources agree that more interdisciplinary studies involving different fields of study are necessary to explore new reliability approaches [14], [15].

1.2 Research material

The data for this thesis are temperature data acquired from temperature measurements of different experiments of a 2-MW outdoor central inverter prototype for large utility-scale PV power plants. The equipment under test (EUT) is different inverter prototypes with varying electronics components from different manufacturers, component configurations and cooling system components. The temperature measurements were performed in a temperature and humidity-controlled laboratory to verify that the implemented EUT corresponds with the simulations during the product development phase. Different filters, power modules and other component configurations were tested to identify the optimal design with high efficiency and competitive pricing, and temperatures in EUT were measured using thermocouples (TC) and thermistors. TC are based on the thermoelectric effect and are relative sensors since they need a temperature reference [16]. Furthermore, thermistors are temperature-sensitive resistors made of ceramic semiconductors that measure temperatures without a temperature reference. Hereafter, a sample refers to one temperature measurement acquired from one component from the whole duration of one test; temperature refers to either surface or junction temperature of an electronic component or any spot inside the inverter; a component refers to any temperature measurement spot from inside the inverter, electrical component or power electronics component; sensor refers to a temperature sensor, thermocouple or thermistor; and load profile refers to the operating point and the ambient conditions of the laboratory.

Figure 2 shows an example of temperature measurements from an EUT under an experiment. Component temperature data are sequential observations taken at discrete times throughout the duration of the test; therefore, it fits the description of time series given in [17]. Even though the observed variable, in this case temperature, is a continuous variable, it is sampled at equal 10 s intervals, and little or no information is lost by measuring the temperature at discrete time intervals [17].

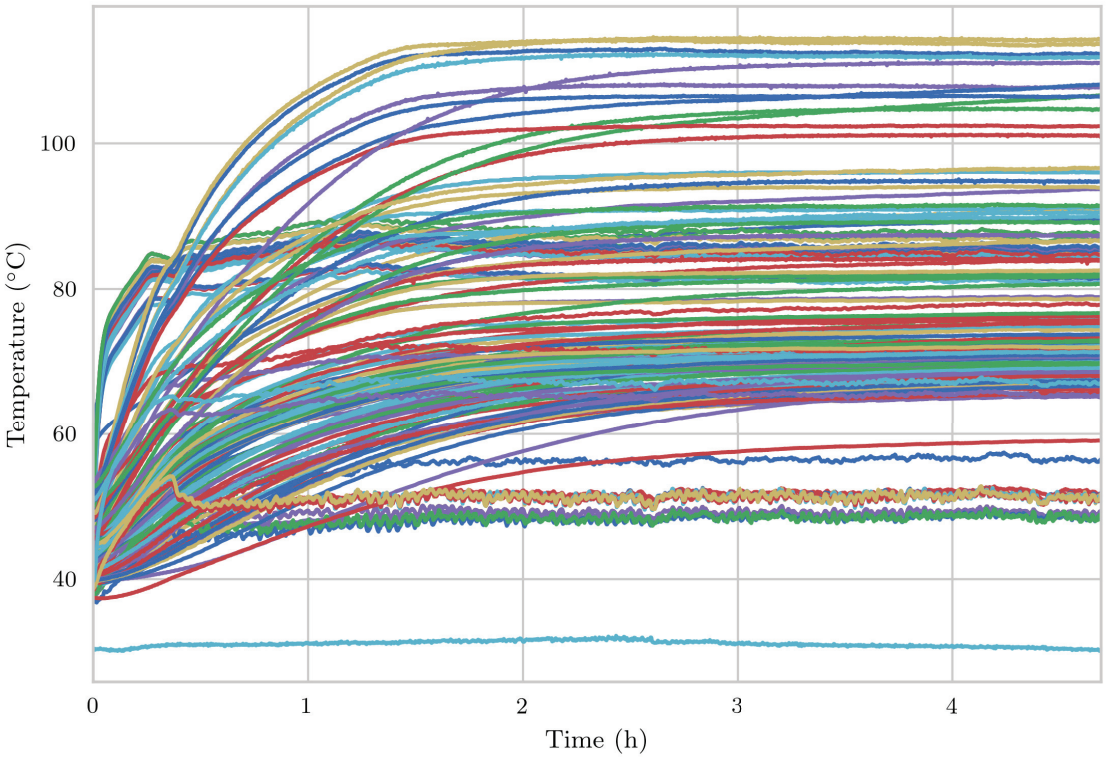


Figure 2: Example of temperatures measured from an experiment during normal inverter operation.

The inverter is an outdoor inverter designed for operating in harsh environments in ambient temperatures ranging between −20 °C and 50 °C [18]. The cooling system is designed to be low maintenance, and the cooling of the system is performed by air and passive heat exchange based on natural convection [19]. The inverter cabinet is also tightly sealed to prevent dust and water from entering the inverter [19]. If the ambient temperature is above 50 °C, the inverter converts less power than the type designed power, while after 60 °C, the inverter stops converting power to the grid. The inverter measures constantly the temperature of electrical components and various other spots inside the inverter. Every component has an optimal operating temperature range, and if the range is surpassed, the inverter converts power at a derated level and eventually stops converting if a maximum temperature threshold is surpassed. To ensure that the components inside the inverter are cooled and

the dissipated heat is transferred out of the system, temperature data are acquired from various spots inside the EUT, including power electronic components' surfaces and junction temperatures, electrical conductors, fuses, switches, breakers, control circuitry, filters, incoming air temperature and the temperature from multiple spots inside the module. Also, the incoming air temperature into the module and the ambient temperature around the module are measured. Moreover, the temperature information is used to determine whether the components can tolerate more current.

The temperature measurements experience disturbances and noise, which Figure 3 shows. Therefore, the measured values are only an estimate of the true temperature value. Noise is classified into inherent noise and interference noise [20]. Inherent noise originates from the sensor or sensor circuitry, while interference noise is transmitted from the background. Moreover, the International Committee for Weight and Measures groups noise uncertainty into two classes, stochastic uncertainty and deterministic uncertainty [21]. Stochastic uncertainty is due to the random nature, while deterministic uncertainty arises from systematic attributes. Sources of the uncertainty shown in Figure 3 include but are not limited to temperature sensor standard error, sensor circuitry, data acquisition system and temperature instability of the laboratory. The data acquisition system and the sensor circuitry are calibrated regularly per the industry standard to maintain their accuracy.

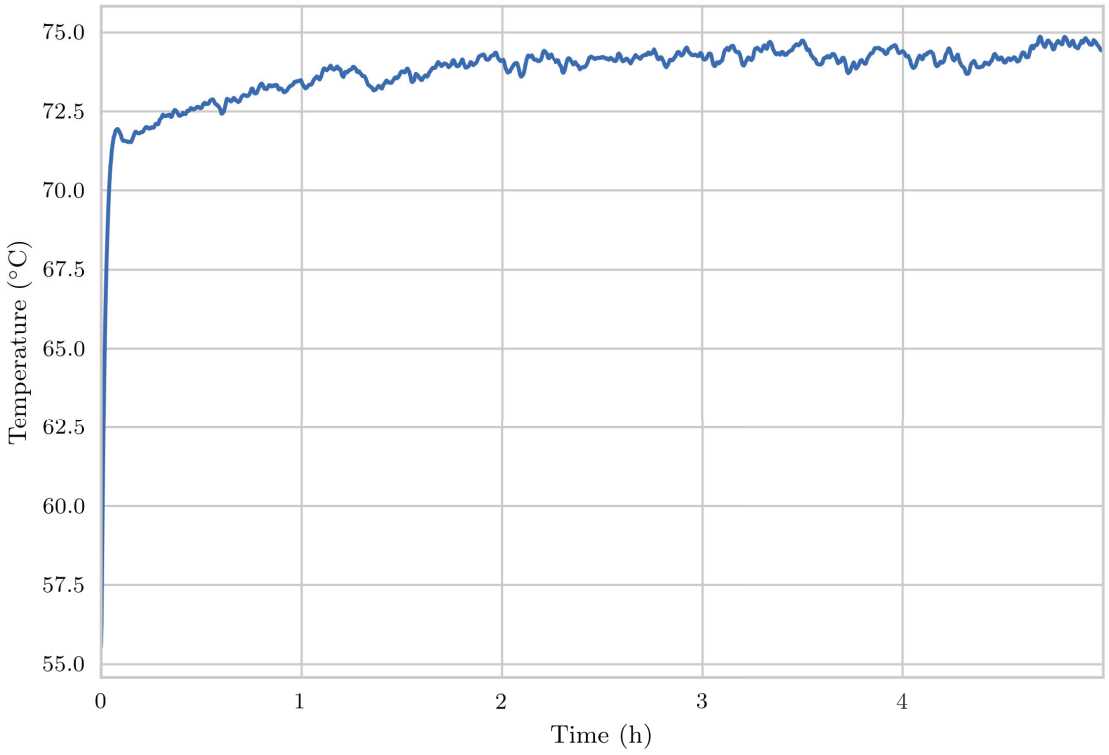


Figure 3: Example of temperature measurement uncertainty from a power electronic component during normal inverter operation.

However, the measurement signal is still distorted due to inherent and interference noise. The inherent noise originating from the temperature sensor makes the measured temperature non-ideal. The inherent noise in the temperature sensors is affected by material quality [20]. While the inherent noise is systematic and within known tolerances, the interference noise arising from temperature instability of the laboratory is irregular and unpredictable. The uncertainty of the system is overcome by incorporating probabilistic models [22].

1.3 Research problems

In the event of temperature sensor failure, the inverter must be shut down, preventing power from being converted to the grid. In a solar energy-based market, loss of power is critical. Exploring the possibility to operate the inverter in case of a temperature sensor failure improves reliability and benefits system owners, PV plant operators and the customers connected to the grid. However, a thorough search of the literature shows that very little research has studied the possibility of continuing inverter operation when a measurement sensor has failed. Moreover, the reliability assessment of operating the inverter without temperature measurements remains untouched.

Therefore, this thesis explores the possibility of estimating the temperature of a single component of an outdoor PV inverter, so the inverter can continue operating without a temperature sensor. On this basis, three research questions are formulated.

1. Can a broken sensor be identified?
2. Can a probabilistic temperature model be formulated for a single measurement for an outdoor inverter?
3. Can the inverter continue operating under the probabilistic model if the sensor is broken?

1.4 Literature review

A frequent argument in the literature is that a grid-connected PV system shares the same attributes as a stochastic process [23], [24], [25], [26]. In most cases, the studied system is not truly random, but stochasticity is used for representing the model uncertainties [27]. Moreover, evaluating the system in a probabilistic manner provides tools to perform risk analysis [22].

1.4.1 Fault-detection

Multivariate analysis (MVA) combines many statistical techniques such as dimensionality reduction and clustering and is considered very useful in systems with multivariate data as an input [28], [29]. Principal component analysis (PCA) is a

dimensionality reduction technique and the use of PCA in an inverter fault-detection context can be found in studies [30], [31], [32], [33] and [34], which all used PCA as a pre-processing technique in the first block in the structure of the fault diagnosis strategy

In [30], dimensionality reduction with PCA showed that PCA improved the fault-detection accuracy of an existing fault-detection model from 85% to 95% [30]. The experimental model in [30] was a physical inverter system, and the fault occurrence was created by removing the power electronic component. Similarly, in [31], PCA was used as a primary-data analysis tool to reduce the size of the input data for a fault detecting model. Then, speed encoder and current sensor faults were simulated in a simulated inverter model. The algorithm detected 98% of the faults in a few seconds [31]. The authors argue that in real-time applications, the significant dimension reduction is crucial to making the algorithm operate quickly by eliminating redundant data [31].

Clustering, or cluster analysis, assesses whether distinct subgroups exist in the data [35]. In relationship to the research problem, the goal is to identify a faulty sensor from a large number of measurements for a large number of components. Additionally, the goal is to identify similar clusters that suggest the components have similar temperature characteristics that can be used later in temperature estimation. Clustering is popular in many fields [28], and numerous cluster analysis methods are listed in [36] such as K-means (KM) and automated algorithms, or machine learning (ML), that make and improve predictions based on previous observations. Also, [28] mentions that KM is the simplest clustering analysis method. Moreover, KM is popular in PV systems output power forecasting and fault-detection studies [37].

Reviewing the literature reveals that cluster analysis methods are not as common in fault-detection applications as in forecasting applications. This is mainly because real-time condition monitoring of power converters is still in its infancy; therefore, its full potential has not been explored [11], [14], [38], [39]. The performances of KM and ML in detecting a fault from time series data are also evaluated in [31]. The systems represent a simulated model of inverter-fed drive in electrical vehicle application. The faults were produced in the speed and current sensors as biases, amplification and loss of signal as well as short circuit faults in the power electronics. According to the results, KM outperformed ML marginally in the implemented real-time fault-detection model [31]. Also, out of nearly 5000 simulation cases, the KM-based clustering analysis detected 98.6% of the failures, while ML identified 97.8% [31].

In data categorization problems, ML proved to be more popular than KM in the recent reviews of [5], [40] and [41]. Also, according to the distribution of different publications with respect to the techniques provided in [5], ML were used in 24% of the reviewed studies, while KM accounted for 6%. However, reasoning for the popularity of ML over other clustering methods was not provided in [5], [40]

or [41]. Nonetheless, reviewing studies utilizing ML, which were [42], [43] and [44], establishes that large number empirical measurement data are required to develop a predictive model with low forecasting error. In the context of this research, the amount of data is significantly smaller, and empirical field data is not available. Therefore, the popularity of ML might not transfer into successful implementation in outlier detection applications, which is the case for this research problem.

Other studies such as [30], [32] and [33] utilized ML as the cluster analysis method in their proposed inverter fault diagnostic systems. The reason for preferring ML over other cluster analysis methods was not given. Moreover, the proposed ML models were trained in an iterative manner by trial and error. The time series data containing few outlier data points representing sensor faults were given as inputs into the algorithm, with the desired output being only the outlier data points. The error between the generated estimates and the actual values is then updated in an iterative fashion until the model detects most of the outliers [30], [32], [33]. Therefore, ML is not applicable to the research problem of detecting an outlier in the data since the quantity of the available data is very small relative to the data used in the studies. The findings in [31] suggest that even simple methods may prove to be accurate and outperform complex models in a fault diagnostic context.

1.4.2 Parameter estimation

Based on the reviewed literature, this is the first time a general methodology is applied to estimate the temperature of an inverter component. However, estimation of the PV system power output is already well established. Since PV inverter temperature estimation is an untouched study, it is worth seeking inspiration from the various models proposed in the literature for estimating the power output of a PV system. The terms estimation and forecasting are used interchangeably.

Thorough reviews of published studies in the field of utility-scale level PV system power output estimation can be found in [5], [40], [41], [45] and [46]. The various forecasting methodologies are divided into four different categories based on their forecast horizon [40], [47]. The forecast horizon categories are very short-term forecasting (VSTF), short-term forecasting (STF), medium-term forecasting (MTF) and long-term load forecasting (LTF) [48] Description of different forecasting horizons and their applications were given in [47].

Very short-term forecasting is essential in real-time monitoring applications [47]. In the context of the thesis problem, very short-term prediction is of interest. If a temperature sensor fault occurs at time t, the temperature of the component at time $t+m$ in the future is estimated, where m is within the range of the very short-term forecast horizon. After step m has been reached, a new estimate is predicted again and the process is repeated in an iterative fashion.

Table 1: Classification of different forecasting horizons [47].

Time step	Application	Forecast horizon
Few seconds to minutes ahead	Real-time monitoring	Very short-term
48 to 72 hours ahead	Unit commitment	Short-term
One week ahead	Maintenance scheduling	Medium-term
Months or years ahead	Network operation planning	Long-term

The very short-term forecasting models can be further classified into stochastic models and ML [48]. The stochastic approach in the field of PV very short-term output power forecasting has proven easy to implement and accurate [45]. However, error metrics assessment in [41] proves ML models outperform stochastic approaches. Similarly, ML models outperformed stochastic models in multiple studies such as in [37] and [49]. However, a large dataset is essential for ML-based models to achieve an accurate representation of the system [47].

[37] used hourly data collected from a 1 MW peak PV plant over the period of two years. In [40], the model was based on minutely recorded data of a full year. Reviewing the summary of publications provided in [46], the smallest dataset in a successful ML implementation is 2 weeks for PV system power output data in the study of [50]. Stochastic methods tend to perform well in both data-poor and data-rich environments, while ML depends on data-rich environments [46]. Data-poor environments refer to the availability of little historical data while data-rich environments refer to the availability of years of observations [46]. The dataset for this thesis accumulatively accounts for 43 hours of operation data under different load profiles. All this indicates that ML is not a viable option to solve the research problem of temperature estimating, given the data available for this thesis.

Sub-classifications of stochastic approaches used in PV power forecasting are Markov processes and time series based models [51]. Markov process is a stochastic model for describing the evolution of a memoryless system and is widely adopted in the probabilistic estimation of the PV system power output [52]. A stochastic process has Markov property if the future value in the system is independent of past values [53]. Consequently, to predict the state of a system in the future, it suffices to consider only its present state and not its history. Markov chain (MC) is important in the application of temperature estimation because it provides both probabilities for different states and a stationary distribution for the process [54]. This information can be used to estimate the temperature and the risk management of continuing operating the inverter when the sensor has failed. Forecasting models based on time series analysis (TSA) have had high degrees of success in PV power forecasting, as argued in [40]. In addition, comparisons between time series analysis and other modelling techniques are presented in [37]. TSA methods are many, but they are all established on the same key properties [26], [45].

The use of MC in forecasting PV system power output is studied in [7], [25], [26],

[55] and [56]. Authors in [7] and [56] proposed an MC-based decision-making model that allows probabilistic scheduling of a PV power generation to meet the demand for power at the right time. In addition, the authors of [25] combined Markov chain with Monte Carlo simulation (MCS) to develop a stochastic model that estimates the generated output power of a PV system. Markov chain Monte Carlo (MCMC) is a method for generating random samples covering many possible outcomes of a given black-box system [5]. In [26] and [55], MC was employed to estimate ST and VST power generation of a grid-connected PV system. In both studies, the model was based on categorized historical data in accordance with weather conditions and operating points of the PV system. With MC, the system was modelled as a series of states, and the transitioning probabilities were calculated. In [26], the error between the 12-h projected estimates and the actual values of a 6-kW PV system power output ranged from 0.4% to 4.8%. In contrast, in [55], the error for 15 min forecast horizon of a 433-kW PV system were at largest 2.47%. In [26], the authors argue that with more historical data, the estimated values become more precise compared to the actual values. Moreover, both [26] and [55] emphasized that successful MC implementation depends on classifying the historical data of the PV system according to their operating points.

Successful implementations of PCA in the field of PV power forecasting can be found in studies such as [42], [43], [57] and [58], in which PCA is not used to forecast per se, but PCA is used to reduce the number of input parameters used in an established model. The one-day ahead (24h) forecasting accuracy of the established model without PCA was 62% at best, as opposed to 80% accuracy with PCA [57]. The authors of [57] also stressed that the results showed that PCA performs better in decreasing the forecast error with higher numbers of PV systems and input variables. Similar findings of PCA improving power prediction performance are confirmed in studies such as [42], [43] and [44]. Both studies [42] and [43] integrated PCA with the established model as a pre-processing procedure to filter noise and improve the power forecasting accuracy. The results in [58] show that the integrated model achieved 95% accuracy with PCA but only 65% without PCA. Also, [43] studied the performance of integrating PCA with the model by analysing the mean absolute error (MAE) and root-mean-square error (RMSE). The MAE and RMSE of the integrated model were both less than 14%, whereas the model without PCA scored 16% for both MAE and RMSE. In a similar fashion, the authors of [44] introduced PCA into an existing model and validated the performance of 24 h forecasting. In addition to improved prediction accuracy, the authors reported reduced computational time of 70% from 1.1 s to 0.3 s compared to an implementation without PCA [44]. Also, the data in [42] and [43] were obtained from 2 kW and 18 kW PV systems, respectively, in contrast to [31], in which the data were obtained from 453 different PV systems with the rated power varying from as low as 4.5 kW up to 750 kW.

Moreover, dimensionality reduction has a positive impact on reducing the amount of required memory for a real-time implementation and diagnosis [34]. In addition,

[59] pointed out that forecasting implementations require a large amount of memory from the embedded system in real-time applications, and the memory-reducing benefits of PCA are stated in [30] and [48], but no numbers are provided.

1.5 Research approach and scope

The intended purposes of the temperature measurements were for verifying the design, cooling systems, firmware compatibility and electrical components and not for the inverter temperature estimation. The quantity and the quality of the available data and the measurement uncertainties present a difficulty. Therefore, the proposed model assumes constant ambient temperature and input power conditions. The components were assumed to remain faultless throughout the operation of the inverter, and a faulty temperature sensor does not indicate a faulty component. Additionally, the cooling system was assumed to be operating flawlessly and not obstructed.

First, initial data analysis (IDA) is conducted to analyse the collected data. Analysing the data with descriptive statistics is an important part of a research process [60]. The importance of IDA in accessing the properties of quantitative data is stressed by multiple works and studies [17], [35], [61], [62], [63]. The data handled in this thesis are time series and examining the time series data with TSA should be conducted before considering any statistical method for system modelling [61]. Therefore, preliminary time series analysis is performed to understand the data-generating process for the purposes of estimation and forecasting.

Rarely is a single approach enough, so combining two or more approaches for the same event that serve distinct purposes inside the system is more beneficial [28], [64]. Based on the research problems, the proposed methodology should consist of two sub-blocks, one for detecting faulty temperature sensors and the second for estimating the temperature and evaluating whether the inverter should continue operating. MVA was applied to the proposed model to reduce the dimension of the data, since the data have hundreds of variables observed over many hours. Furthermore, MVA was applied to assess whether discovering patterns between the measurement variables and subgroups among the observations is possible. Then, MC was used for temperature estimation, since there is a strong theoretical basis for using it to model the inverter system and it has been used successfully in many PV system power output forecasting applications. Probabilistic modelling does not require knowledge of the internal system to model it [27]. Moreover, when data are limited and uncertainties between the variables are present, probabilistic forecasting methods are more advantageous than other forecasting methods [48].

1.6 Structure of the research

This thesis is structured as follows. Section 2 establishes the essential concepts in mathematics and statistics that are used throughout the thesis. Statistical evaluation of the data is conducted in Section 3. Furthermore, Section 3 established the frameworks for the research problem and presents the proposed model. Section 4 provides and discusses the results obtained from the proposed model. The conclusion and recommendations for future work are given in Section 5.

2 Theoretical foundations

2.1 Mean and median

The mean and the median are measures of location [65]. Although both measure the central location for a distribution, there are differences between these two location measures. The population mean μ of discrete sample space X is given by the following:

$$\mu = \mathrm{E}[X]\,. \tag{1}$$

An estimate for the population mean is the arithmetic mean $\bar{x}$. For a data set containing observations $x_1, x_2, \ldots, x_n$, the arithmetic mean $\bar{x}$ is defined by the formula in Equation (2) [65]:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{x_1 + x_2 + \cdots + x_n}{n}\,. \tag{2}$$

Median is the middle value when the observations are ranked in order of magnitude [60]. For example, when considering the set of temperature observations in order of magnitude in Equation (3), the median is the middle value, which is 72 °C.

$$\{70\,^{\circ}\mathrm{C}, 70\,^{\circ}\mathrm{C}, 72\,^{\circ}\mathrm{C}, 72\,^{\circ}\mathrm{C}, 73\,^{\circ}\mathrm{C}\}\,. \tag{3}$$

2.1.1 Robustness and breakdown point

The introduction of outliers in the observations affects the summation term significantly in Equation (2). A large outlier in contrast to the rest of the observations makes the mean a less reliable measure of central tendency. However, the median is not affected that easily by outliers. The median in Equation (3) remains the same after introducing an outlier observation. For a finite set of an odd number of observations, the breakdown point of the median can be derived as follows. Let the observations be in order of magnitude $x_1, x_2, \ldots, x_n$. Then, the median can be derived as follows:

$$\{x_1, x_2, x_3, \ldots, x_{\frac{n+1}{2}-1}, x_{\frac{n+1}{2}}, x_{\frac{n+1}{2}+1}, \ldots, x_{n-2}, x_{n-1}, x_n\}\,. \tag{4}$$

In the sample space of n observations, the median is the middle observation $x_{\frac{n+1}{2}}$. The median point divides the sample space into two equal parts that have $\frac{n+1}{2} - 1$ number of samples. That is, the left side of the equation $x_1, \ldots, x_{\frac{n+1}{2}-1}$ has in total $\frac{n+1}{2} - 1$ observations, just as the right side of the equation $x_{\frac{n+1}{2}+1}, \ldots, x_n$ has in total $\frac{n+1}{2} - 1$ observations. Either one of the $\frac{n+1}{2} - 1$ samples can be contaminated with outliers and still that would not affect the median. For finite sample-size n the breakdown point is given by the formula in Equation (5). Table 2 summarizes the breakdown points for the sample mean and median as $n \to \infty$ [66].

$$\left\lfloor \frac{n-1}{2n} \right\rfloor\,. \tag{5}$$

Table 2: Summary of breakdown points [66].

Estimator	breakdown point
Sample mean	0
Sample median	0.5

As n grows larger, the asymptotic breakdown point of the median becomes $\frac{1}{2}$, meaning that up to half of the observations can be contaminated without affecting the median. Therefore, the median is less susceptible to outliers and is a reliable measure of a central tendency with very noisy data. However, in the absence of outliers, the mean is the better measure of central tendency than the median [65].

2.2 Dispersion measures

Standard deviation (SD) and variance are measures of statistical dispersion. Let $\mu = \mathrm{E}[X]$ be the expected value of the variable X, then the variance of X is the following:

$$\mathrm{Var}(X) = \mathrm{E}\left[(X-\mu)^2\right] , \tag{6}$$

and the standard deviation of X is:

$$\mathrm{SD}(X) = \sqrt{\mathrm{Var}(X)} . \tag{7}$$

The variance of a discrete set of observations n, which is equally likely, can be written as the following:

$$\mathrm{Var}(X) = \frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2 . \tag{8}$$

The variance measures the spread of the distribution. $\mathrm{Var}(X) = 0$ indicates no deviation in the observations, i.e. the observations are identical. In contrast, larger variances correspond to more deviated observations. In that sense, the variance is a measure of risk and represents uncertainty in the observed variable.

2.3 Independence measures

Covariance and correlation coefficients are measures of statistical dependence measuring the linear relationship between two variables [35]. The difference between covariance and correlation is that covariance is unit dependent because it keeps the units of the variables X and Y. Therefore, no meaning can be interpreted from comparing covariances of different variables. Correlation coefficient normalizes the variables on a -1 to 1 scale. This normalization of data removes the units of the

variables so the correlation coefficient becomes interpretable between different variables that vary in units.

Covariance is an essential measure of estimation and forecasting and multivariate analysis since most descriptions and formulas are expressed in terms of covariance [61]. Because of its essential role in virtually everything, it is important to explain covariance in more depth.

2.3.1 Covariance

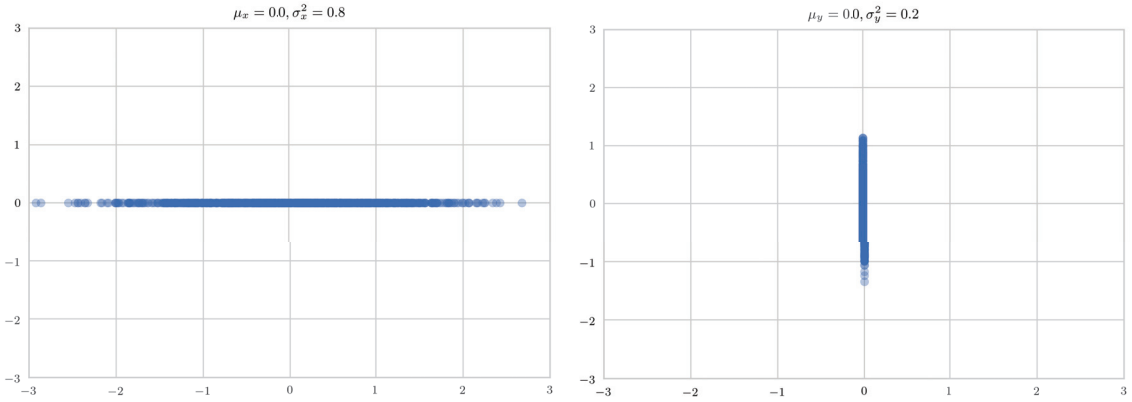
The covariance $\sigma(x, y)$ or $\mathrm{Cov}(x, y)$ is defined as follows [35]:

$$\sigma(x, y) = \mathrm{E}[(x - \mathrm{E}(x))(y - \mathrm{E}(y))]\,. \tag{9}$$

Variables X and Y are centred by subtracting their corresponding mean. Centred scores are then multiplied to measure whether a change in one variable is associated with a change in the other variable. Lastly, the expected value of the centred scores product is calculated. Similarly, the covariance $\sigma(x, x)$ or $\mathrm{Cov}(x, x)$ is obtained by:

$$\sigma(x, x) = \mathrm{E}[(x - \mathrm{E}(x))(x - \mathrm{E}(x))]\,. \tag{10}$$

The $\sigma(x, x)$ is also known as variance, or $\mathrm{Var}(X)$, and describes the scattering of the data in the directions parallel to the axes of the variable space. The projection of the sample data onto a one-dimensional space is shown in Figure 4:



(a) Variable X. (b) Variable Y.

Figure 4: Distribution of 1000 samples for variables X and Y drawn from $\mathcal{N}(0,\, 0.8)$ and $\mathcal{N}(0,\, 0.2)$.

For these samples, variance $\sigma(x, x)$ clearly explains the spread the horizontal spread along the x-axis, and variance $\sigma(y, y)$ explains the vertical spread along the y-axis. Next, the same samples in 2D space are shown in Figure 5.

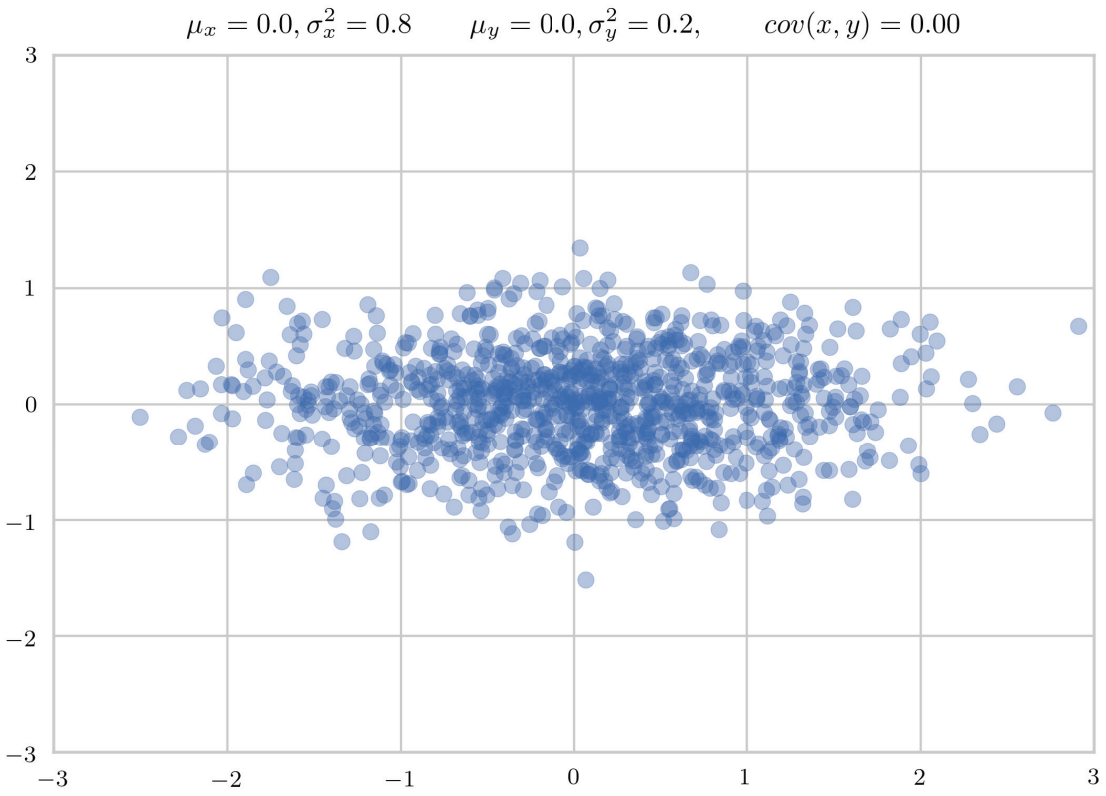


Figure 5: Distribution of X and Y samples in 2D space.

Although the samples are scattered in a 2D space, the effect of the variance on the spread of the data is still evident. In particular, the spread is strongest along the horizontal axis. Also, the cov(X, Y) is zero, meaning the random variables are uncorrelated. The figure shows that the increase in x-values does not increase the y-values. Instead, there are random scatterings along the vertical axis.

For 2D data, the $\sigma(x, x)$, $\sigma(y, y)$, $\sigma(x, y)$ and $\sigma(y, x)$ can be represented by a covariance matrix Σ, or also known as a variance-covariance matrix [67]:

$$\Sigma = \begin{bmatrix} \sigma(x,x) & \sigma(x,y) \\ \sigma(y,x) & \sigma(y,y) \end{bmatrix}. \tag{11}$$

The main diagonal of the matrix has the variance of the variables X and Y, while the entries outside the main diagonal are the covariances. Two-dimensional data can be represented by a 2×2 covariance matrix. Likewise, the spread of three-dimensional data is represented by a 3×3 covariance matrix, and the spread of N-dimensional data by an $N \times N$ covariance matrix.

The covariance matrix for the 2D data represented in Figure 5 is therefore the

following:

$$\Sigma = \begin{bmatrix} 0.8 & 0 \\ 0 & 0.2 \end{bmatrix}. \tag{12}$$

Consider there is a positive correlation between the variables X and Y and let $\text{Cov}(X, Y) = 0.15$. Then, the covariance matrix is the following:

$$\Sigma = \begin{bmatrix} 0.8 & 0.15 \\ 0.15 & 0.2 \end{bmatrix}. \tag{13}$$

The entries outside the main diagonal of the covariance matrix are nonzero since the variables are now correlated, and Figure 6 shows the effect of the covariance on the scatterplot. After the orientation of the data has changed, there is a clear diagonal correlation explained by the introduction of the covariance. Figure 6 shows that the x-value increases, on average, the y-value also increases because of the positive covariance. Therefore, the horizontal and vertical spreads of the data are explained by the variance. In contrast, the orientation of the data spread is explained by the covariance.

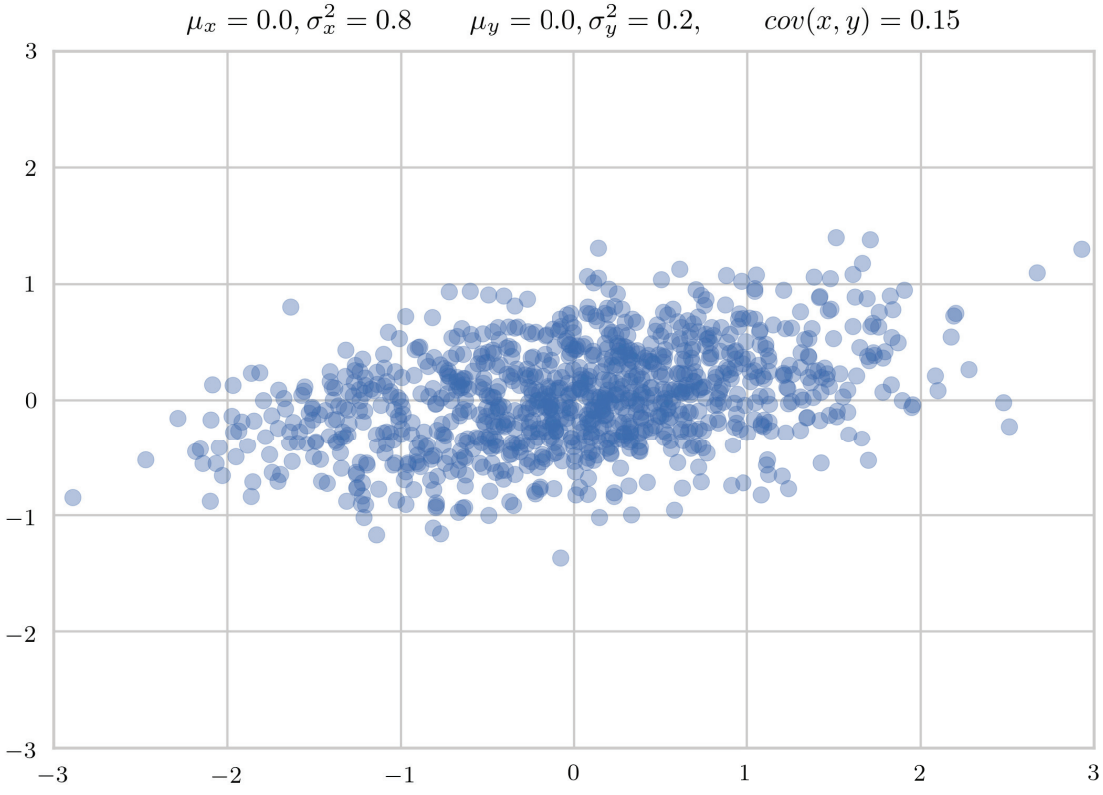


Figure 6: Distribution of X and Y samples in 2D space.

The study of matrices is closely related to linear algebra [67]. As Figures 5 and

6 show, variance defines the spread of the data, and covariance defines the orientation of the data. Therefore, the covariance matrix represents a linear operator that transforms the shape of the data [67]. An intuitive way to characterize the information that the covariance matrix portrays could be vector algebra. Furthermore, the direction, magnitude and orientation of the spread can be represented with a vector in the direction and with a magnitude that equals the of the scattering of the data. The vector characterization is related to the eigendecomposition of a covariance matrix [67]. Therefore, it is important to define the eigenvectors and eigenvalues of the covariance matrix.

There are several ways to compute the eigenvectors and eigenvalues [67], and the most common approach is to find for a square $N \times N$ matrix $\mathbf{A}$, a scalar λ and a nonzero vector $\mathbf{u}$ that satisfy Equation (14) [68]:

$$\mathbf{A}\mathbf{u} = \lambda\mathbf{u}\,. \tag{14}$$

Equation (14) can be rewritten as [68]:

$$(\mathbf{A} - \lambda\mathbf{I})\mathbf{u} = 0\,, \tag{15}$$

where λ is a scalar called the eigenvalue associated to the eigenvector $\mathbf{u}$ and $\mathbf{I}$ is the $N \times N$ identity matrix. To obtain nontrivial solutions that are not zero, Equation (15) is set to the following to find the λ values [68]:

$$|\mathbf{A} - \lambda\mathbf{I}| = 0\,. \tag{16}$$

Equation (16) is known as the characteristic equation [68]. Solving the characteristic equation for λ yields nontrivial values for λ that can be later substituted into Equation (15) to find corresponding values of $\mathbf{u}$.

According to the linear algebra theorems, the eigendecomposition of any symmetric matrix $\mathbf{A}$ can be written as the following [67]:

$$\mathbf{A}\mathbf{U} = \mathbf{U}\mathbf{\Lambda}\,, \tag{17}$$

or as the following:

$$\mathbf{A} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^{-1}\,, \tag{18}$$

where each column of $\mathbf{U}$ is an eigenvector of $\mathbf{A}$ and the diagonal elements of $\mathbf{\Lambda}$ give the eigenvalues.

Therefore, the eigenvectors and eigenvalues of the covariance matrix define the linear transformation of the covariance matrix as the following:

$$\mathbf{\Sigma}\mathbf{u} = \lambda\mathbf{u}\,, \tag{19}$$

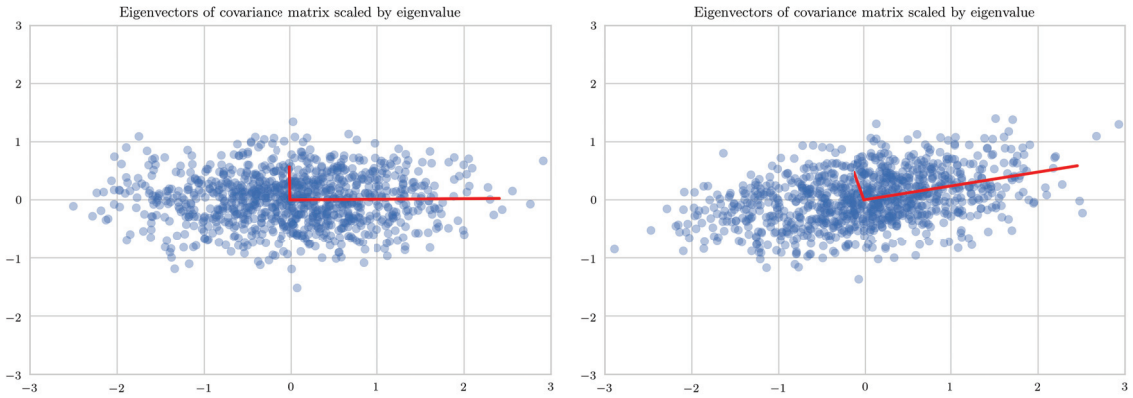
or also represented as the following:

$$\mathbf{\Sigma} = \mathbf{V}\mathbf{D}\mathbf{V}^{-1}, \tag{20}$$

where $\mathbf{u}$ is the eigenvector of Σ and λ is the corresponding eigenvalue. Then, applying the eigendecomposition on the covariance matrix in Equation (12) via Equations (14) and (15) yields to the following:

$$\begin{bmatrix} 0.8 & 0 \\ 0 & 0.2 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 0.8 & 0 \\ 0 & 0.2 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}^{-1}. \tag{21}$$

The eigenvalues are $\lambda_1 = 0.8$, $\lambda_2 = 0.2$ and eigenvectors $\mathbf{u}_1 = (1, 0)$, $\mathbf{u}_2 = (0, 1)$ for the covariance matrix in Equation (12). The implication is that the variances are equal to the eigenvalues λ, and the eigenvectors $\mathbf{u}$ point to the direction of the spread in the data. Figure 7 shows the magnitude and the direction of the eigenvectors in the two cases.



(a) Covariance matrix in Equation (12). (b) Covariance matrix in Equation (13).

Figure 7: Eigendecomposition of covariance matrices.

The eigendecomposition of the two-dimensional data resulted in two eigenvector components, and eigenvectors are often referred to as principal components (PC) [69]. Similarly, 3×3 three-dimensional data can be represented by three eigenvector components. Consequently, n-number of eigenvector components capture the spread of N-dimensional data. However, unlike the two- or three-dimensioned data, N-dimensional data cannot be plotted. Nonetheless, because eigendecomposition captures the essential information on the spread and variance of the data, eigendecomposition of the covariance matrix is the basis of many machine learning (ML), dimension reduction and cluster analysis techniques and time series analysis methods [36], [68], [69], [61].

These examples establish that the covariance matrix can be utilized to transform high dimensional data into lower dimensional space [69]. In Equation (20), $\mathbf{V}$ is the orthogonal matrix, and $\mathbf{D}$ is the diagonal scaling matrix [67], [68], [69]. The orthogonal refers to the rotation and diagonal to the coordinate-wise scaling and can be used to project the data matrix to a new feature space with different orientation and scale.

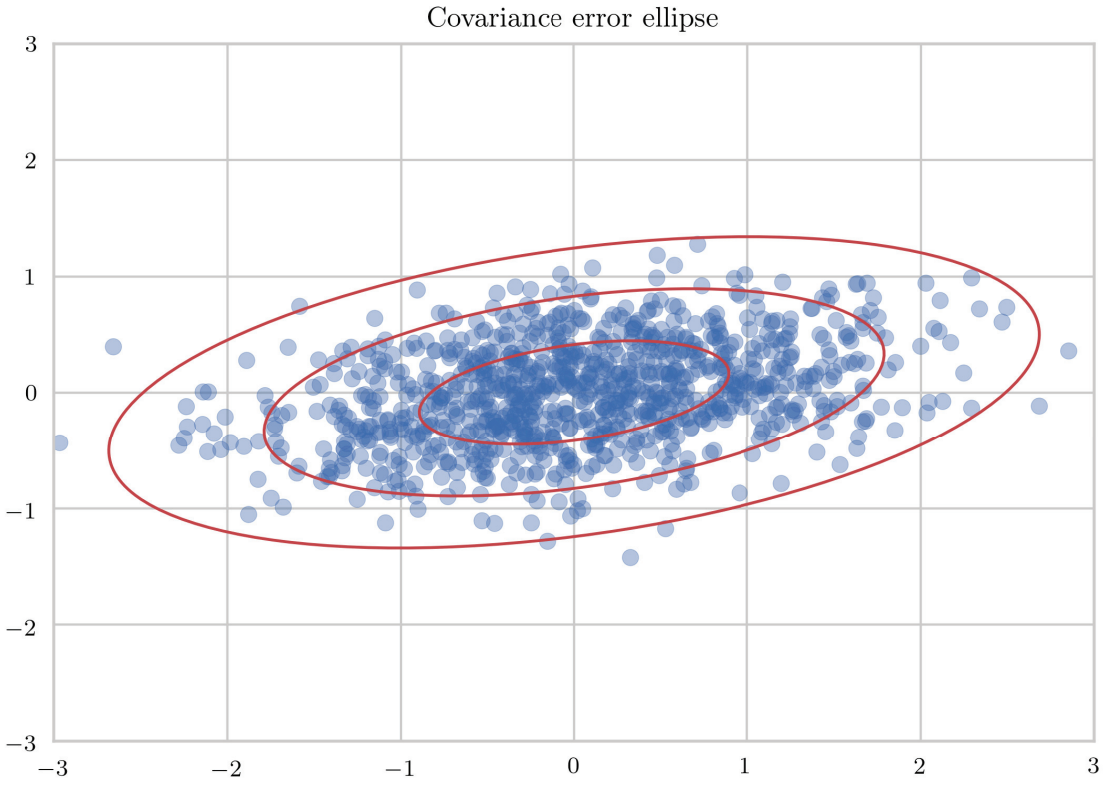


Figure 8: Covariance ellipses containing 99%, 95% and 68% of the data.

Another interpretation or application of the eigenvectors is to construct a confidence set such as an ellipse [35]. In the case of two-dimensional data, the eigenvectors can be used to measure the length and height of the ellipse around the scatterplot [60], and the ellipse can be used to set the boundaries, capturing certain probability density of the data. The variances of the variables are used to scale the ellipse to capture certain probability density from the data [67]. The ellipse defined by the eigenvalues λ_1, λ_2 and eigenvectors $\mathbf{u}_1$, $\mathbf{u}_2$ of the covariance matrix is formed by the following linear combination [67]:

$$\sqrt{\lambda_1 \cos(t)\mathbf{u}_1} + \sqrt{\lambda_2 \sin(t)\mathbf{u}_2}\,, \tag{22}$$

for t in the interval $[0, 2\pi]$. Figure 8 shows contour plots for the ellipse of the covariance matrix in Equation (13), and each ellipse contains a different proportion of observations from the data.

2.3.2 Correlation coefficient

A correlation coefficient is a statistical measure of the strength of relationships between two variables [64], and there are various methods for calculating the correlation coefficient. The three of the most widely used methods are Kendall tau, Spearman rank correlation and Pearson correlation [60]. Each approach measures the degree of correlation on a -1 to 1 scale, but they vary in the type of strength of association measure. Among the various correlation coefficient techniques, the Pearson correlation coefficient is the most commonly used in the literature [64]. In either case, a correlation coefficient of 0 means no correlation exists between the changes in the two variables. As the correlation coefficient approaches ± 1, the correlation becomes stronger. A positive correlation coefficient sign means that an increase in one variable increases another variable. In contrast, a negative correlation coefficient sign means that a decrease in one variable decreases the other variable.

3 Research design and methodology

3.1 Statistical evaluation of data

The time series data of a selected group of components are plotted in Figure 9, while the corresponding frequency distribution, a histogram, is illustrated in Figure 10. The histogram suggests that some components have similar temperature distributions, which can be seen by the overlapping histograms. Figure 11 illustrates the temperature distribution of one selected component across different tests. The figure shows that some tests have similar load profile characteristics, while the other tests differ significantly.

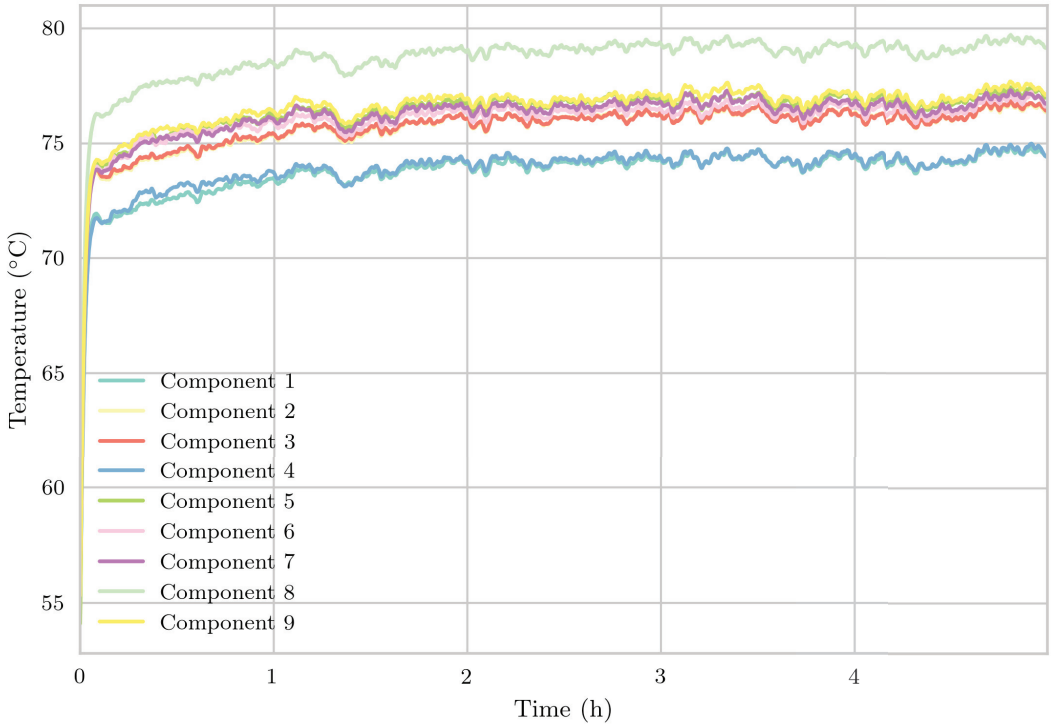


Figure 9: Temperature time series of a selected group of components in one experiment.

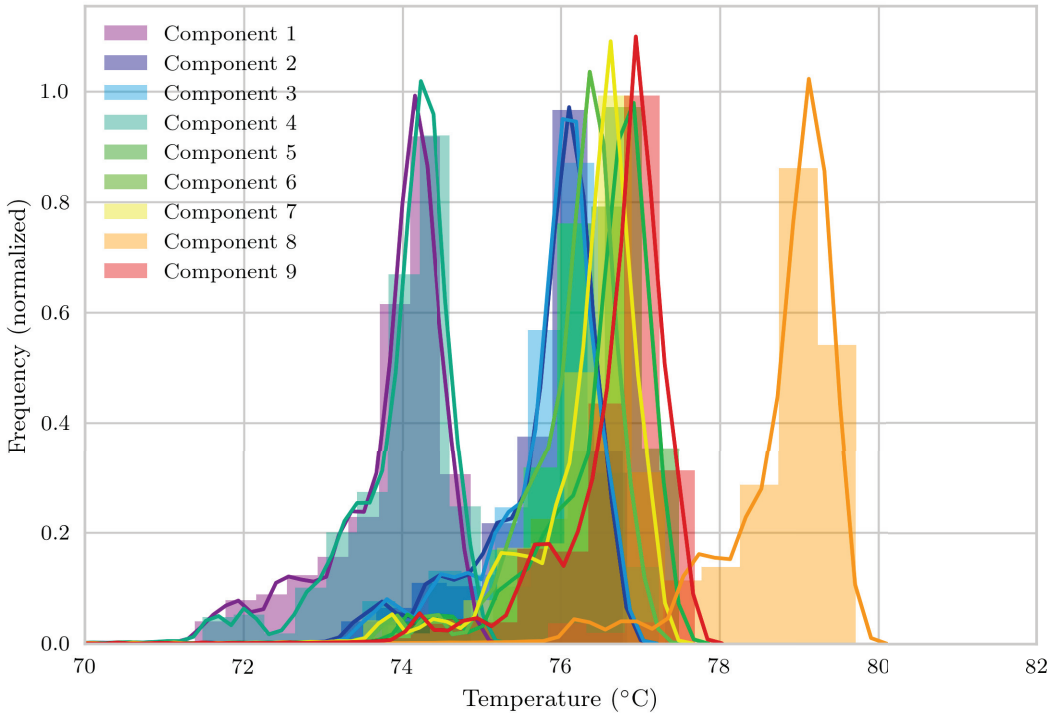


Figure 10: Temperature distribution of a selected group of components in one experiment.

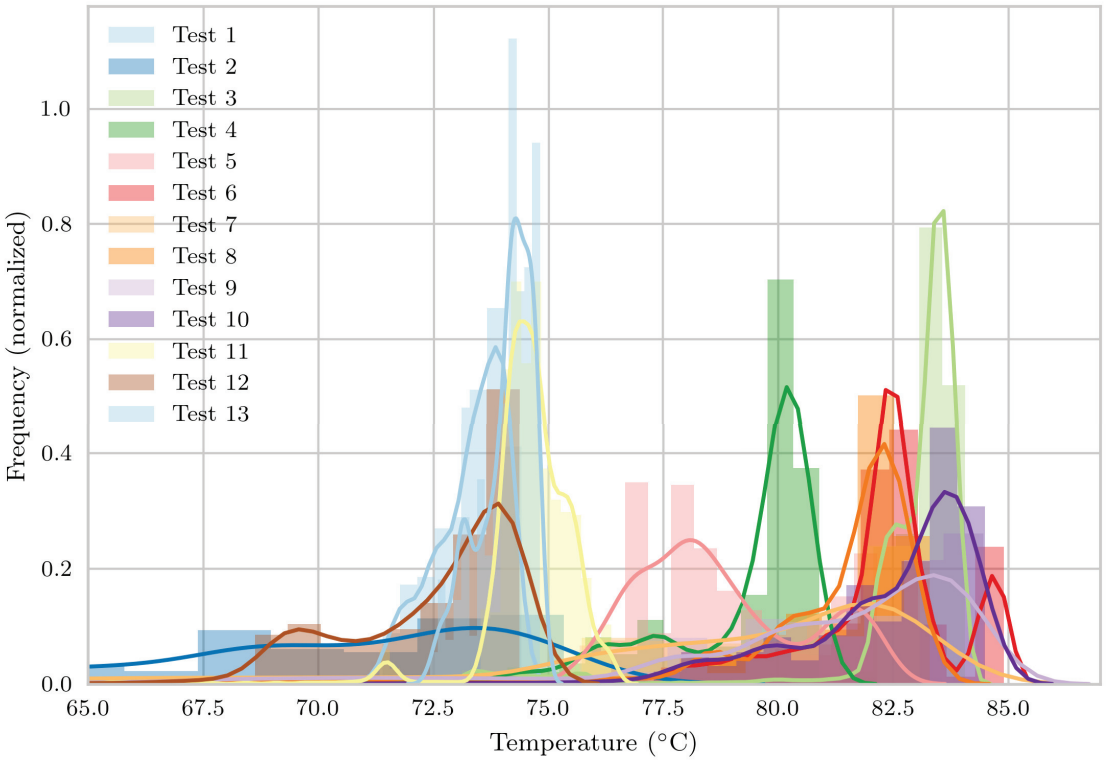


Figure 11: Temperature distribution of one component across multiple experiments.

3.1.1 Independence measures

Figure 12 presents the similarity matrix of various components and shows a strong relationship between all 9 components. However, this relationship is not entirely true since Figure 9 shows that components 4, 5 and 8 have different temperature profiles from the remaining components. Moreover, the temperatures of components 4 and 5 are similar, while component 8 has a significantly higher temperature. The component temperatures do indeed rise over time, so a strong positive correlation is expected. Therefore, correlation is not a versatile measure of the relationship between the temperatures of components in this study.

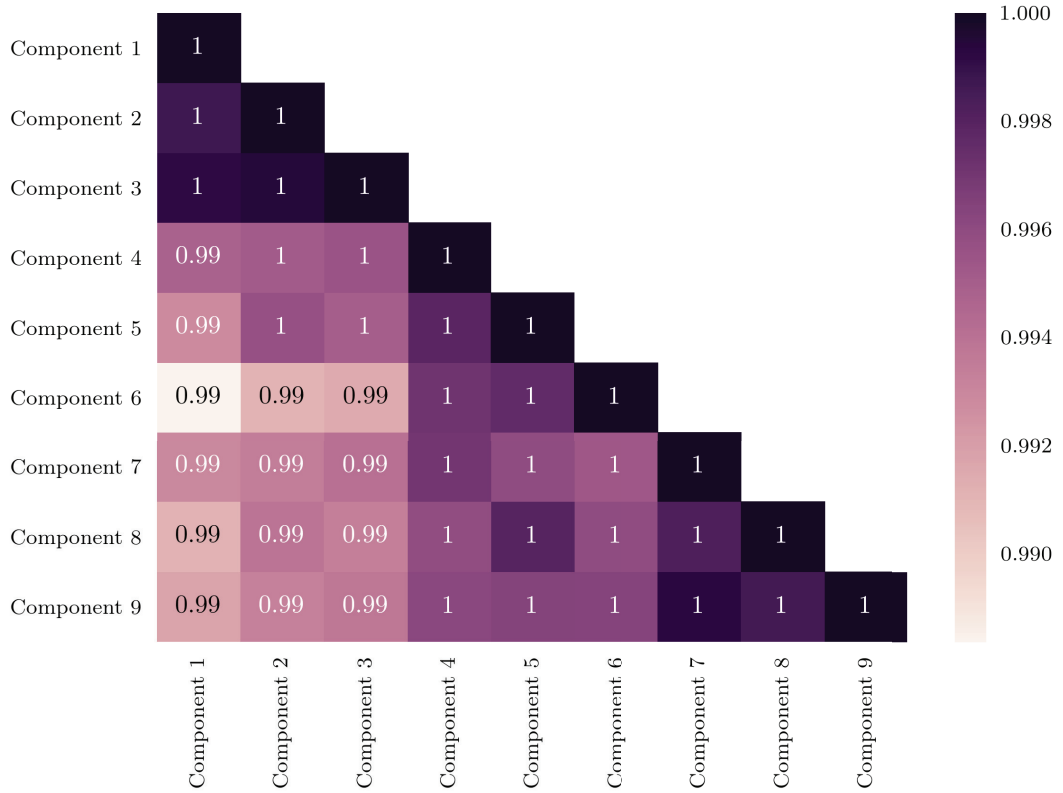


Figure 12: Similarity matrix constructed with the Pearson correlation coefficient of a selected group of components from one experiment.

3.1.2 Time series analysis

Time series is considered a combination of trend (T_t), seasonality (S_t) and noise components (N_t) [17]. The trend is defined as the increasing or decreasing value in the series. The oscillatory behaviour in the series indicates a seasonal component, and the random variation or non-systematic component in the series is the noise. Therefore, time series analysis involves decomposing the three aforementioned components [61]. Naturally, all time series data cannot be represented with only these

three components [17]. If the time series data x_t were to consist of systematic patterns, x_t can be written as the sum of the components:

$$x_t = T_t + S_t + N_t \,, \tag{23}$$

or as the following:

$$x_t = T_t \cdot S_t \cdot N_t \,. \tag{24}$$

Equation (23) is additive model, and Equation (24) is multiplicative model [61]. To illustrate TSA, the solar power generation data from March to July 2018 was considered, as shown in Figure 13 obtained from the Finnish national electricity transmission grid [70].

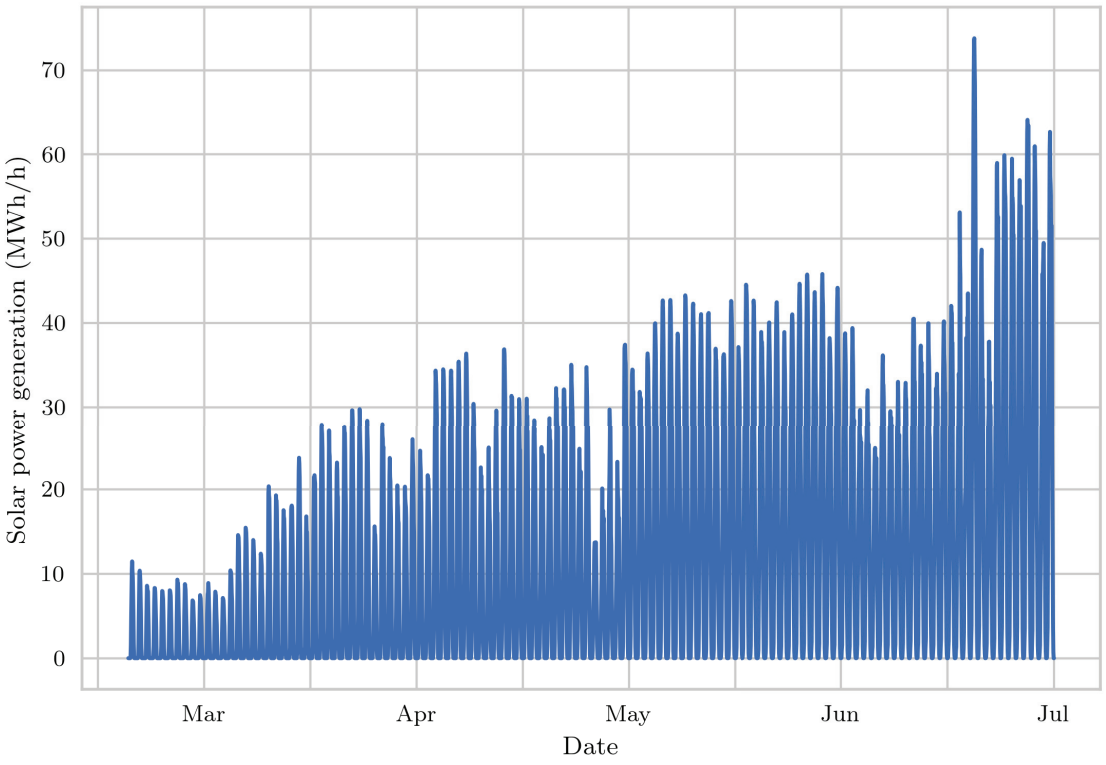


Figure 13: Solar power generation in Finland from March 2018 to July 2018 [70].

In Figure 14, time series analysis was performed, and the three components were decomposed using modules provided by Statsmodels [71]. Analysing the data begins by observing the dominant patterns. The results show a non-linear upward trend as well as a seasonal component that changes within the day in a cyclic pattern. The results reveal that the solar power generation increases with the season and that there are variations within the day.

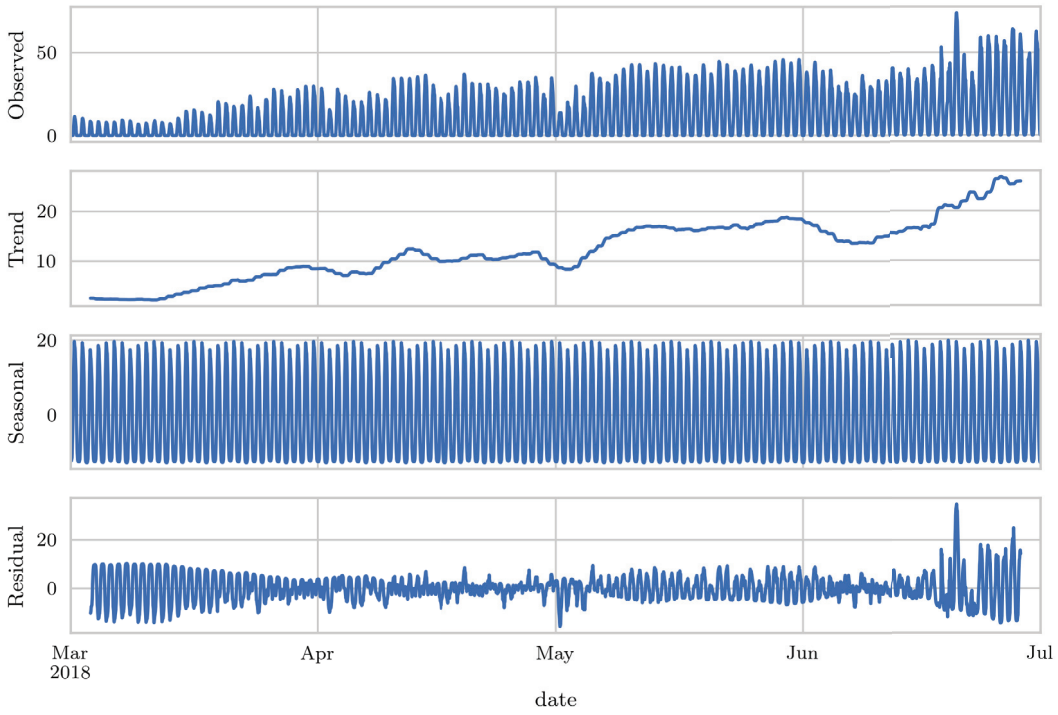


Figure 14: Time series analysis of generated solar power [70].

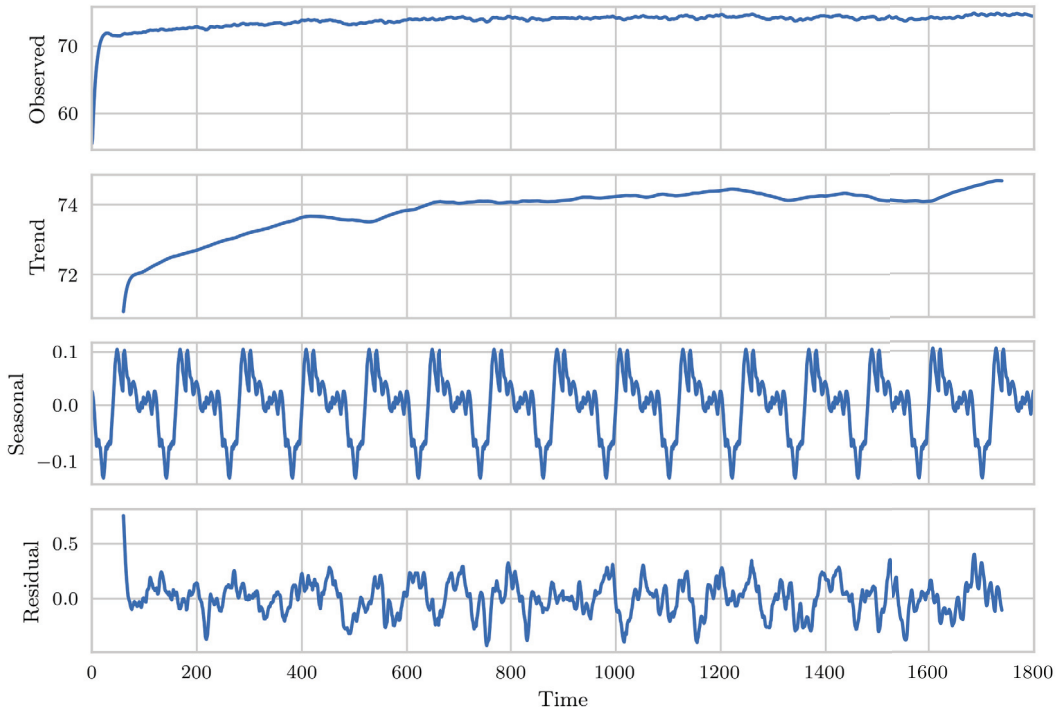


Figure 15: Time series analysis of the selected component temperature data.

A similar analysis was applied to the temperature data presented in Figure 3, and Figure 15 shows the results of the analysis. There are no trend factors; the cyclic nature is due to the poor laboratory ambient temperature control scheme, while the residuals are due to the measurement noise. However, two key characteristics can be extracted from Figure 15. The temperature does stabilize at some point to certain temperatures, after which there are no trends. This is the foundation for solving the problems of the thesis. The time of sensor failure relative to the stabilization time is the fundamental question. If the sensor failure occurs after the temperature stabilization, the temperature stays stable given that the cooling system is not obstructed and the operating point remains the same. However, if a sensor failure occurs before the stabilization time, then the system needs to estimate the stabilization time and evaluate the probability of reaching the threshold value in a future time instant. Since the stabilization time is critical, the noise must be filtered to extract an accurate stabilization time.

3.2 Model formulation

Figure 16 illustrates a black-box system representation of an inverter, and power converter systems can be modelled in terms of input power P_{in}, output power P_{out} and power loss P_{loss} [1]. The input P_{in} is the DC power from the PV panels. The system output response variables are the P_{loss} and the AC power P_{out}. Combined together, P_{loss} and P_{out} are the system output response to the input variable P_{in}. In many resources, the inverter losses P_{loss} are considered to be purely heat dissipation [14], [38].

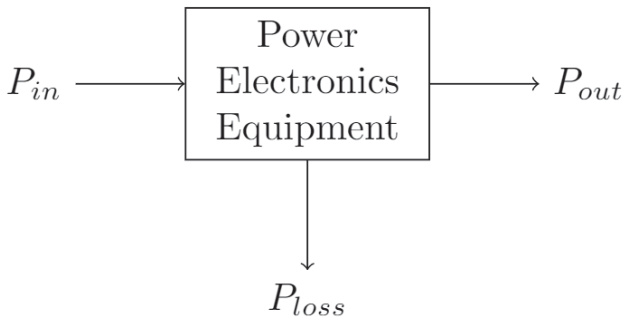


Figure 16: Power converter system in terms of P_{in}, P_{out} and P_{loss} [1].

Figure 17 illustrates the outdoor inverter as a system in terms of input and output variables, and the outdoor inverter takes many user input parameters before converting power to the grid. Some of the user input parameters are the output reference reactive power $P_{g,ref}$, output reference reactive power $Q_{g,ref}$, the DC voltage of the PV panels U_{dc}, desired AC voltage U_g, grid voltage variation $U_{g,var}$ and grid frequency f_g. The vector sum of active and reactive power gives apparent power

S, which is obtained from the formula $S = \sqrt{P^2 + Q^2}$. The response variable of interest is the response output variable, the thermal dissipation P_{loss}.

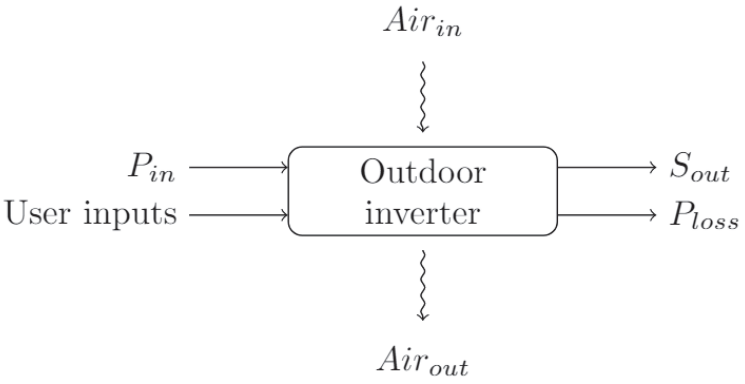


Figure 17: Inverter system model.

3.3 Detecting a faulty temperature sensor

Each test contains m number of temperature observations of n number of components $X_1, X_2, \ldots, X_n$. The response variable of interest Y detects a failed temperature sensor, and the variable is discrete and has two states: the sensor is faultless or faulty. The temperature sensor outputs are always a numerical value, not the status of its health. Even a faulty temperature sensor outputs a signal that has a numerical value interpretation. In that sense, detecting a faulty sensor is challenging. However, a faulty temperature sensor exhibits abnormal behaviour compared to the other faultless sensors. Therefore, the goal is to analyse the observations of variables $X_1, X_2, \ldots, X_n$, to find out whether the observations fall into relatively distinct subgroups: normal and outlier data. This kind of approach for discovering hidden structures in data and putting data into groups is known as clustering [28], which is a special case of multivariate statistical inference, or multivariate analysis (MVA) [35]. Clustering finds homogeneous subgroups among the observed data, and dimensionality reduction finds a low-dimensional representation of the observed data that captures most of the variance [28].

3.3.1 Dimensionality reduction

Let n be the number of observed components, variables, within a test, and the number of measured samples taken from the whole duration of the test in discrete time is m. The dimension of the collected data is $m \times n$, and it is advantageous to arrange the variables and the observations in an $m \times n$ matrix $\mathbf{X}$ called the data matrix [72]:

$$\mathbf{X} = \overbrace{m \text{ observations}}^{n \text{ variables}} \left\{ \begin{pmatrix} x_{1,1} & x_{1,2} & x_{1,3} & \cdots & x_{1,n} \\ x_{2,1} & x_{2,2} & x_{2,3} & \cdots & x_{2,n} \\ x_{3,1} & x_{3,2} & x_{3,3} & \cdots & x_{3,n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{m,1} & x_{m,2} & x_{m,3} & \cdots & x_{m,n} \end{pmatrix} \right. ,$$

where the rows correspond to the observation time and the columns to the components, or variables. Thus, the temperature sample of component number 1 measured at time instant 2 is stored in the second row of the first column.

Finding the covariance or correlation coefficient between each variable in the temperature measurements in a set of n variables, $X_1, X_2, \ldots, X_n$, as part of the initial data analysis results in $\binom{N}{2}$ different combinations. With $n = 300$, there are 44850 different correlation coefficients. Then, with large n values, it is not possible to analyse all the correlation coefficients in an insightful way and derive meaning from them. Some of the information in large data might be redundant [5], and a more appropriate approach is required for analysing the data when the n and m are large. Ideally, the optimal result would be to find a low-dimensional representation of the data without redundant information.

PCA is a very common dimensionality reduction method used with multivariate data [67], [28], and PCA is utilized in PV inverter research in [31], [57], [73], [30], [32], [42], [58] and [43]. The essence of PCA follows the theory explained in Chapter 2. PCA transfers a $m \times n$ data to a lower dimension while retaining most of the information of the original set [67]. In practice, the vector $\mathbf{x} \in \mathbb{R}^{m \times n}$ is transformed into $\mathbf{z} = \mathbf{R}^T\mathbf{x}$, with $\mathbf{z} \in \mathbb{R}^{k \times n}$ where $k \ll m$. The transformed variables k are called principal components (PC) [67]. However, there are multiple approaches to apply PCA to extract the PC vectors [69]. The PC vectors can be extracted through singular value decomposition (SVD) of the original data matrix or with eigenvalue decomposition of data covariance matrix [31]. As shown in Chapter 2, the geometric representation and the eigendecomposition of the covariance are intuitive, but most PCA implementations perform SVD for the eigenvalue problem since it is computationally efficient.

Regardless of the way of decomposing the eigenvectors and eigenvalues, the computed eigenvectors are sorted in an order of magnitude according to the corresponding eigenvalues. The eigenvector with the largest eigenvalue is the first PC vector, the second largest eigenvector is the second PC vector and so forth. Cumulatively, the sum of all principal components explains 100% of the original data. The number of principal components in the new subspace $\mathbb{R}^{k \times n}$ can be reduced by selecting a number of principal components, k, that captures most of the variance. The re-

duction of dimensionality keeps most of the information and therefore is a good approximation of the original data [31]. The general steps for performing PCA are listed in the following list, and a detailed description has already been established in Chapter 2.

- The data are normalized.
- $m \times n$ dimensional covariance matrix is computed.
- Eigenvectors and corresponding eigenvalues are calculated.
- Eigenvectors are arranged in order of magnitude by the corresponding eigenvalues.
- k number of eigenvectors with the largest eigenvalues are selected to form a $k \times n$ dimensional matrix.
- The data are transformed onto the new $k \times n$ subspace

Across the reviewed studies, such as [57] and [58], the selection of k number of PC vectors was done so the k number of selected PC vectors cumulatively explained at least 85% of the variance in the original data.

3.3.2 Clustering

KM clustering seeks to divide the observations into a designated number of clusters [28]. For the KM algorithm to work, the observations must be normalized. In this study, the observations are temperatures in Celsius, so normalization is not needed. However, unlike other methods, the KM algorithm cannot be presented in equations [29]. Instead, a detailed description of the algorithm progression is given as in [36].

Let $C_1, \ldots, C_k$ denote the centre, or centroid, of each cluster. Given an integer k and set of observations $X \in \mathbb{R}^{m \times n}$, the KM algorithm first generates centroids $C_1, \ldots, C_k$ and places the centroids in random locations in the $\mathbb{R}^{m \times n}$ space [29]. After placing the centroids in random locations, the following is repeated iteratively until convergence [36].

- For each observation x_i:
    1. find the nearest centroid C_j, so that the distance between x_i and C_j is minimized;
    2. assign the observation x_i to cluster j;
- For each cluster $centroid_j$:
    1. calculate the mean of observations inside the cluster; and
    2. assign the mean as a new position for the centroid.

Each observation x_i is assigned to a cluster to which the centroid is closest. After assigning each observation to a cluster in the first iteration, the centroid values are recalculated by taking the arithmetic mean of all observations in the cluster. Then, after the new position for each centroid is found, the algorithm performs the same iteration until convergence has been reached. Furthermore, when the newly calculated position for the centroids no longer changes, an optimum has been reached. The distance metric depends on the type of data, but Euclidean distance is often used [28], and pre-assigning the number of clusters might sound counter-intuitive. However, in a fault-detection context, a two-cluster division is intuitive and the sensor data is either normal or outlier.

3.4 Estimating temperature

3.4.1 System description as a stochastic process

A stochastic process is a process or a system that evolves randomly in time [74]. If the system is observed at a set of discrete times, it is a discrete-time stochastic process. Then, if the system is observed continuously, it is a continuous-time stochastic process [53]. In this thesis, the system was observed at a set of discrete times, and time was considered a subset of the non-negative integers $0, 1, 2, ..., n$.

Consider the temperature measurements of a component in the inverter plotted in Figure 3 as a system. The system is observed at times $t = 0s, 10s, 20s, 30s, ...$, and the system can be considered to evolve randomly in time. Let the temperature measurement observed at any given time t be X_t. The set of random variables $S = \{X_0, X_1, X_2, X_3, ..., X_t\}$ are then the states of the system, or the state space of the stochastic process. The stochastic process in this case is written as $X_t, t \geq 0$ and the observed temperature values of the system are realizations of the stochastic process [53].

Stochastic modelling gives tools for predicting the temperature of the component at a future time instant $t + m$ or for computing the probability of reaching a temperature threshold [53]. The change of the observed temperature of the component was considered a result of random events. In other words, given the current state of the system X_t, any other information about the past is irrelevant for predicting the future state X_{t+m}. The process has no memory; in addition, the probability to move from one state to another is random and only depends on the current state, meeting the characteristics of Markov chain (MC) [54]. Also, stochastic processes in which the state at time $t + m$ is determined by the state at time t and not by the states before t are called Markov processes [74].

3.4.2 Markov chain

Consider the mentioned discrete-time stochastic process X_t, $t = 0, 1, 2, 3, ...$, where the system is a single component and the states of the system are the measured

temperature values. The temperature at time t, X_t, takes values in a finite set S, and all the possible values for X_t are the states of the system. Suppose the system is observed at times $t = 0, 1, ..., 9$ and yields the temperature observations $X_0, X_1, ..., X_9$. The question concerning the thesis problem is whether it is possible to predict the state of the system at time $t = 10$. To answer such a question, first, the transition probabilities of the system must be established for every n and every finite sequence of observations $(i_0, ..., i_n)$ [65]:

$$P\big\{X_n = i_n | X_0 = i_0, ..., X_{n-1} = i_{n-1}\big\}, \quad i = 1, .., N\,. \tag{25}$$

A transition probability is the likelihood of being in state j given the process started at state i. If the transition probabilities are established, with conditional probability, X_n depends on the earlier states $X_0, X_1, ..., X_{n-1}$, which can be written as the following:

$$P\big\{X_n = i_n | X_0 = i_0, ..., X_{n-1} = i_{n-1}\big\}\,. \tag{26}$$

However, since the temperature measurement of the component is assumed to be Markov process, it satisfies the Markov property. In other words, the next state depends only upon the current state making the knowledge of the past states redundant. This knowledge simplifies the equation, which can be written as the following:

$$P\big\{X_n = i_n | X_0 = i_0, ..., X_{n-1} = i_{n-1}\big\} = P\big\{X_n = i_n | X_{n-1} = i_{n-1}\big\}\,. \tag{27}$$

Since the system uses discrete time, it is called a discrete-time Markov chain (DTMC). Note that the system is indeed observed at discrete times, but the time and temperature variables are continuous. A continuous-time Markov chain could be used, but little or no information is lost by modelling the system as DTMC [17]. A DTMC $X_n, n \geq 0$ is time homogeneous if, for all $n = 0, 1, 2, ...,$

$$P(X_{n+1} = j | X_n = i) = P(X_1 = j | X_0 = i)\,. \tag{28}$$

Thus, the one-step transition probability depends on i and j, but the probability is the same at all times n and hence does not depend on the time n. For time homogeneous DTMCs, the one-step transition probability is the following:

$$p_{i,j} = P(X_{n+1} = j | X_n = i), \quad i, j = 1, 2, ..., N\,. \tag{29}$$

If the system consists of many states, it is advantageous to construct an $N \times N$ matrix $\mathbf{P}$ called the probability matrix, transition matrix or stochastic matrix [65]:

$$\mathbf{P} = \begin{pmatrix} p_{1,1} & p_{1,2} & p_{1,3} & \cdots & p_{1,N} \\ p_{2,1} & p_{2,2} & p_{2,3} & \cdots & p_{2,N} \\ p_{3,1} & p_{3,2} & p_{3,3} & \cdots & p_{3,N} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ p_{N,1} & p_{N,2} & p_{N,3} & \cdots & p_{N,N} \end{pmatrix}\,.$$

The one-step probability of transitioning from state 1 to state 2 is stored in the first row of the second column, that is $p_{1,2}$. However, when constructing a transition matrix from sample data, it must satisfy the properties in Equations (30) and (31) [53].

$$0 \leq p_{i,j} \leq 1, \quad 1 \leq i, j \leq N\,. \tag{30}$$

$$\sum_{j=1}^{N} p_{i,j} = 1, \quad 1 \leq i \leq N\,. \tag{31}$$

That is, each transition probability is non-negative and between zero and one. Since each row of a transition matrix represents all the possible outcomes given the current state, the sum of all probabilities is indeed one.

Consider that the system has three different temperatures, $S = T_0, T_1, T_2$, that make the states of the system. For example, at time n, temperature T_1 is observed. The system is observed again at time $n + 1$. There are three possible outcomes: either the system is still at T_1 or at T_2 or T_0. Consequently, the state can either repeat itself or move from one state to other. Figure 18 shows the transition diagram for this DTMC. The transition from any state to another is possible. For example, a move from T_0 to T_2 is possible, and such a case is not plotted in the figure for the sake of simplicity.

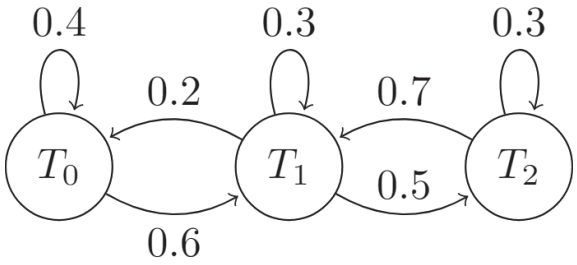


Figure 18: Example of a state transition diagram between states in Markov chain.

Probability $p_{0,0} = 0.4$ refers to the probability that the observation T_0 is observed again after one time increment. Probability $p_{0,1} = 0.60$ is the probability of measuring the temperature T_1 if the previous observation was T_0. Each state can repeat itself or move from one state to another. Similar representation can also be made of its transition matrix. With the transition probabilities given in Figure 18, the transition probability matrix $\mathbf{P}$ takes the following form:

$$\mathbf{P} = \begin{pmatrix} 0.4 & 0.6 & 0.0 \\ 0.2 & 0.3 & 0.5 \\ 0.0 & 0.7 & 0.3 \end{pmatrix}\,.$$

The transition probability matrix $\mathbf{P}^1$ gives the transition probabilities at $t = 1$, $\mathbf{P}^2$ gives the transition probabilities at $t = 2$, and so forth [74]. In this study, the transition probabilities are calculated from the time series data for each component, and the time series data are sequences of different temperature measurements indexed by time. The sample space is comprised of the measured temperatures rounded to the nearest decimal. Then, the probabilities are calculated by mapping each state to another state in an iterative one-step transition manner:

- For each state S_i:
    1. find all occurrences of states S_i in the data;
    2. for each S_i occurrence, determine:
        (a) does the state repeat itself?;
        (b) the state before the transition to S_i;
        (c) the state after the transition from S_i; and
    3. compute the transition probability for S_i.

At the end, a transition matrix is obtained that satisfies the properties in Equations (30) and (31).

Considering the problem of this thesis, MC offers many interesting properties and includes stationary distribution. As $n \to \infty$, the transition probability matrix $\mathbf{P}^n$ converges to a probability distribution known as stationary distribution [65]. In other words, the transition probability matrix remains unchanged after n time periods. Then, the stationary distribution π is solved by Equation (32) [74]:

$$\pi = \pi \mathbf{P}\,, \tag{32}$$

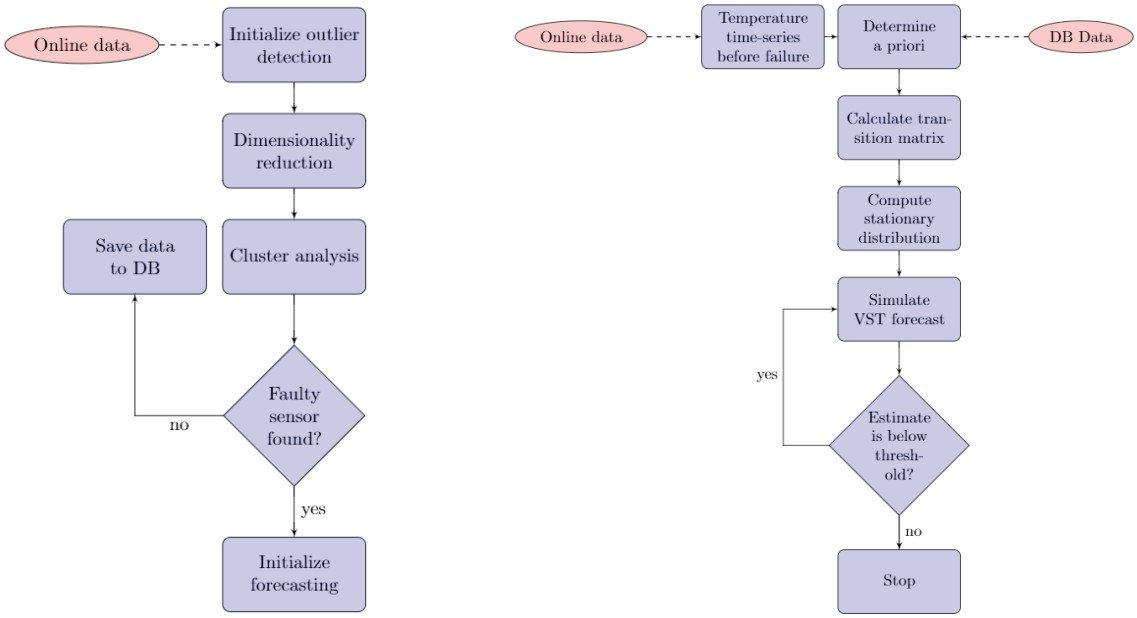
satisfying the property,

$$\sum_{j=1}^{N} \pi_{i,j} = 1, \quad 1 \leq i \leq N\,. \tag{33}$$

The utility of solving the stationary distribution in this thesis is modelling the MC steady-state probabilities. The stabilization time of each component is not known, so it is reasonable to model the MC of each component for time t_n, $n \to \infty$ and compute the stationary distribution of each component.

3.5 Proposed model

Figure 19 shows a block diagram of the proposed model. The inputs provided to the model are real-time temperature time series data of each component. The proposed model analyses the data, saves it to a database (DB) and detects a faulty temperature sensor. The model also consists of two sub-blocks, an outlier detection block and a forecasting block.



(a) Outlier detection block.

(b) Forecast block.

Figure 19: Flowchart of the proposed model.

3.5.1 Outlier detection

Figure 19a shows the progress of the outlier detection model. Large sets of time series data are inputted into the model. In the first step, the outlier detection algorithm processes the data into a data matrix to later generate an $m \times n$ covariance matrix and applies PCA. PCA is implemented via the eigenvalue decomposition of the covariance matrix to obtain the eigenvectors and the corresponding eigenvalues. Finally, the least numbers of the eigenvectors are selected as PC vectors so the corresponding eigenvalues are as large as possible. The new reduced space dimension is defined by the number of PC vectors. The k number of PC vectors was selected to cumulatively explain at least 85% of the variance in the original data since this was agreed upon in various studies [57], [58].

The inspiration for outlier detection was derived from the error ellipse discussed in Chapter 2, which is also known as density ellipse [68]. First, the model draws an error ellipse or an ellipsoidal using the empirical covariance obtained from the observations. The objective is to maximize the density of the ellipse by capturing the maximum number of points inside the envelope. Then, the model compares the density of the error ellipse to the already established error ellipses of the training data saved in the DB. However, even the error ellipse can be over-fitted to contain outliers [72]. This is overcome by introducing a cost function, as suggested in [75]. Including an observation far from the rest of the observations is penalized since it decreases the density of the error ellipse. The process is iterative and is repeated throughout the flow of input data. After the inverter operation has ended, the error ellipse is saved in the DB.

The data are classified into different operating point categories, clusters, with the KM algorithm and using Euclidean distance as a distance metric. The degree of similarity between the observations is measured with the distance of each centroid. Hypothetically, centroids that are close to each other represent similar operating conditions. For example, consider a temperature sensor failure for component x_i, which belongs to cluster C_j. By predetermining which clusters are similar to C_j, the temperature data of x_i from similar conditions can be obtained to estimate the current temperature.

If an observation is deemed outlier, the time series data of the component are saved to the DB but are classified as outliers. The forecasting block is then initialized to forecast the temperature of the component. The outlier detection block passes on the following information:

- time series data of the observed component until the moment of failure;
- stabilization time and temperature of the component derived from DB; and
- a priori distribution of time series data derived from DB.

A priori distributions reflect the beliefs before seeing any data [76]. When the temperature of a component cannot be observed due to a failed sensor, assumptions about the possible temperature tendency can still be made. The purpose of using a priori distribution is to predict future temperature data, given past beliefs about the temperature of the component.

3.5.2 Extracting stabilization time and temperature

The time series data are filtered from noise to obtain accurate estimates of the time of stabilization and stabilization temperature. Specifically, the stabilization time is given as the first time instant when the temperature changes over an observation period of a half hour; it is 0.5 °C, i.e. $\Delta T = 0.5$ °C during $\Delta t = 30$ min [77]. Median filter was used due to its robustness, as has been established in Chapter 2. In a median filter, a sample was obtained within a predefined window, and the median was calculated. Later, the window was shifted by one time increment, and the process was repeated [61]. Figures 20 and 21 illustrate the robustness of a median filter. Also, the temperature measurement presented in Figure 3 was first filtered with the median filter. Afterwards, the steady state was computed from the filtered data. A sample window of 30 minutes was used, and the change of temperature ΔT was calculated. Then, the window was later shifted by 10 seconds, and the ΔT was calculated again until the stabilization time was found.

Figure 20 shows the steady state was reached after approximately 47 minutes, and the stability temperature was 73 °C. The original data were then contaminated with noise sampled from the normal distribution $\mathcal{N}(0, 1)$ so only 51% of the data

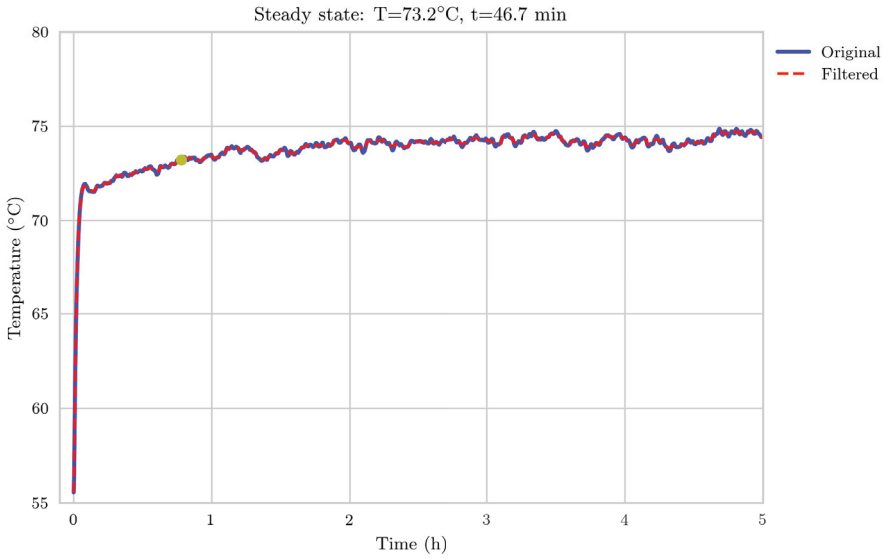


Figure 20: Original data filtered with median filter.

were original data and 49% of the data were noise. Then, filtering the contaminated data reveals that the steady state temperature was achieved at 47 minutes and the stability temperature was 73 °C. This was nearly identical with the steady state time and temperature computed with the uncontaminated data, thereby illustrating the robustness of the median filter.

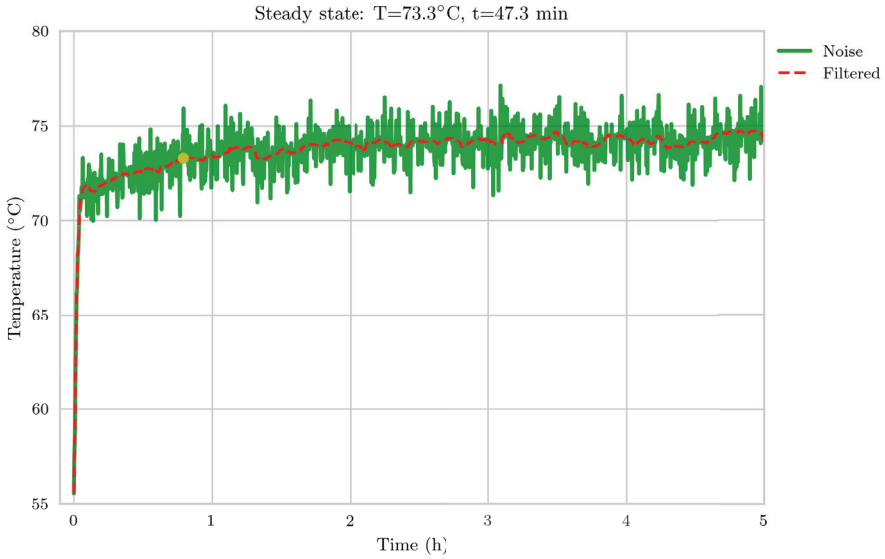


Figure 21: Contaminated data filtered with a median filter.

3.5.3 Saving data to the database

Data are filtered with a median filter, and the stabilization time and temperature are computed and saved. Additionally, a similarity matrix is established between the observations from different tests. When sensor fault occurs, a similarity matrix is constructed between previously saved test data and all the observed time series data collected until the moment of failure. After similar test data have been computed, a priori data can be established. A priori data are previously recorded time series data of the component with a faulty sensor. If the similarity between the current test and the previous test is high, the a priori data should be an accurate estimate of the current temperature time series.

3.5.4 Temperature estimation

The relationships established in the outlier detection block are used to simulate temperature time series in future time instances. The metric used to measure the relationship between two variables is the Euclidean distance measured from lower dimensional projection and computed by the empirical covariance matrix. Instead of trying to estimate precise temperature values at a future time instant $t + m$, hundreds of time series are simulated from the MC state matrix to analyse the distribution of the temperature among all forecasted time series. When analysing the distributions, the emphasis is placed on when the system reached steady state, i.e. determining the stability time and stability temperature. If the projected stability temperature is higher than the temperature threshold, then the run is aborted.

When the temperature sensor fails at time t, the last temperature measurement is saved as the initial state. A priori data of the component is passed on from the outlier detection block, which in turn retrieves the data from the database after establishing the proximity matrix of similar historical data. MC steady state matrix is computed and time series are then simulated, with the last saved temperature being the starting point of each simulated time series. The distribution of the temperature is analysed among all simulated time series. Simulating hundreds of time series consider the uncertainty related to forecasting the temperature at a future time instant $t + m$. Projecting different scenarios is a robust approach for the decision-making process of continuing the inverter operation [51].

3.6 Performance metrics

For robust statistics, root-mean-square error (RMSE), median (M), and median absolute deviation (MAD) were used as performance metrics. Median was then used to measure the central location of the distribution and the MAD to measure the dispersion of the distribution. Afterwards, RMSE was used to calculate the difference between the actual values and the model estimated values. The RMSE and

MAD are defined by the formulas in Equation (34) and Equation (35) respectively.

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - x_{est})^2}\,. \tag{34}$$

$$\text{MAD} = \text{M}\left(\,|x_i - \text{M}(x_{est})|\,\right)\,. \tag{35}$$

4 Results

To examine the accuracy of the proposed model, a series of simulation experiments were performed. The operation of the outlier detection block was validated first, and then the whole proposed model was validated. The data for the validation are divided into two categories: training data and validation data. The division is done as 70% training data and 30% validation data since that division is commonly used in the literature to verify the performance of a model [47]. Data from experiments 1–10 were the training data while data from experiments 11–14 were the validation data. Experiments 11–14 were selected as the validation data since they contain outlier data and make 30% of the available data. Three case studies are conducted: 1) baseline test, 2) outlier detection test and 3) temperature estimation. The first test is a baseline case with training data that includes no outlier data to verify the ability to block to the expected standards. The system aims to classify the data into correct groups by clustering the data and finding the centroid of each cluster. Moreover, a baseline is established for the error ellipse from the validation data. Since the cluster analysis and outlier detection algorithm are based on geometry, the dimensionality reduction is critical.

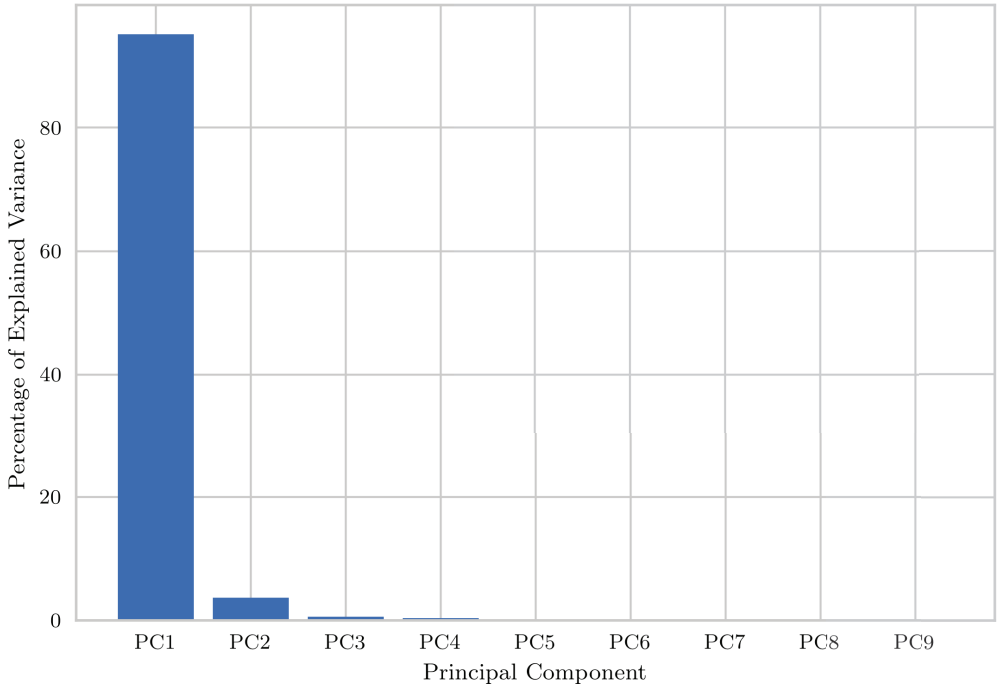


Figure 22: Explained variance ratio per each principal component in data without outliers.

4.1 Baseline test

Data from 14 different tests without outliers were inputted into the model separately. First, the dimension of the test data was reduced, and the centroid of each cluster was then calculated by the KM algorithm to use later to find similar clusters. Figure 22 shows the results of PCA. The histogram shows the percentage of explained variance per each principal component. The explained variances of the first two components are 95% and 4%. Cumulatively, they explain 99% of the variance. Therefore, projecting the results of the cluster analysis into two dimensions does not cause information loss. However, for meaningful visual representation, the results of the PCA and cluster analysis are presented in three dimensions, which the axes are not proportional in magnitude. PC1 explains 99.3% of the variance, while PC2 explains 0.4% of it. Therefore, one step in the PC1 direction is more significant than in the PC2 direction.

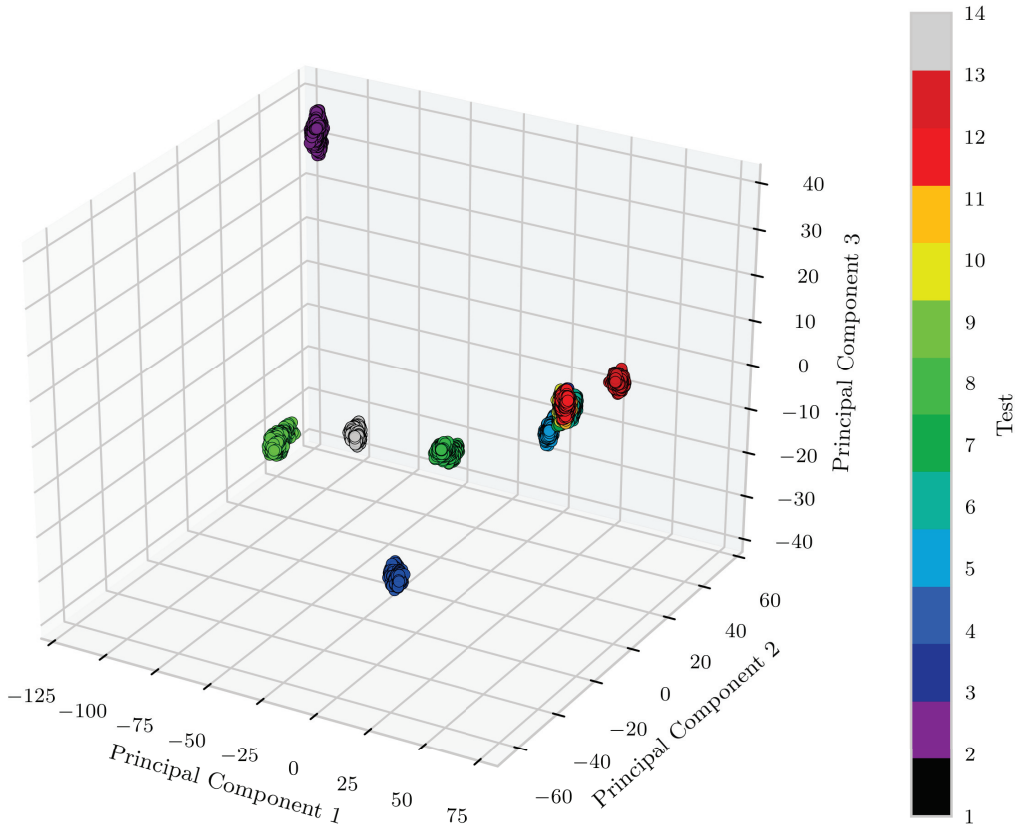


Figure 23: Clustering of the observations from various tests in 3D.

Cluster analysis in Figure 23 shows that some tests overlap while others are far apart. Constructing a similarity matrix based on the Euclidean distance from the centroid of each cluster in Figure 24 reveals more information about which tests are more similar. The Euclidean distances between each test are calculated in three-dimensional space. A lower Euclidean distance indicates greater resemblance

between the tests. The proposed model suggests validation tests 14, 13, 12 and 11 are similar to tests 1, 8, 6 and 7 based on the Euclidean distance of each centroid after the dimension reduction, which is built on the mean and variance of each observation. To verify the suggestion, the root-squared errors of the temperatures of identical components from each test pair are calculated and compared in Table 3.

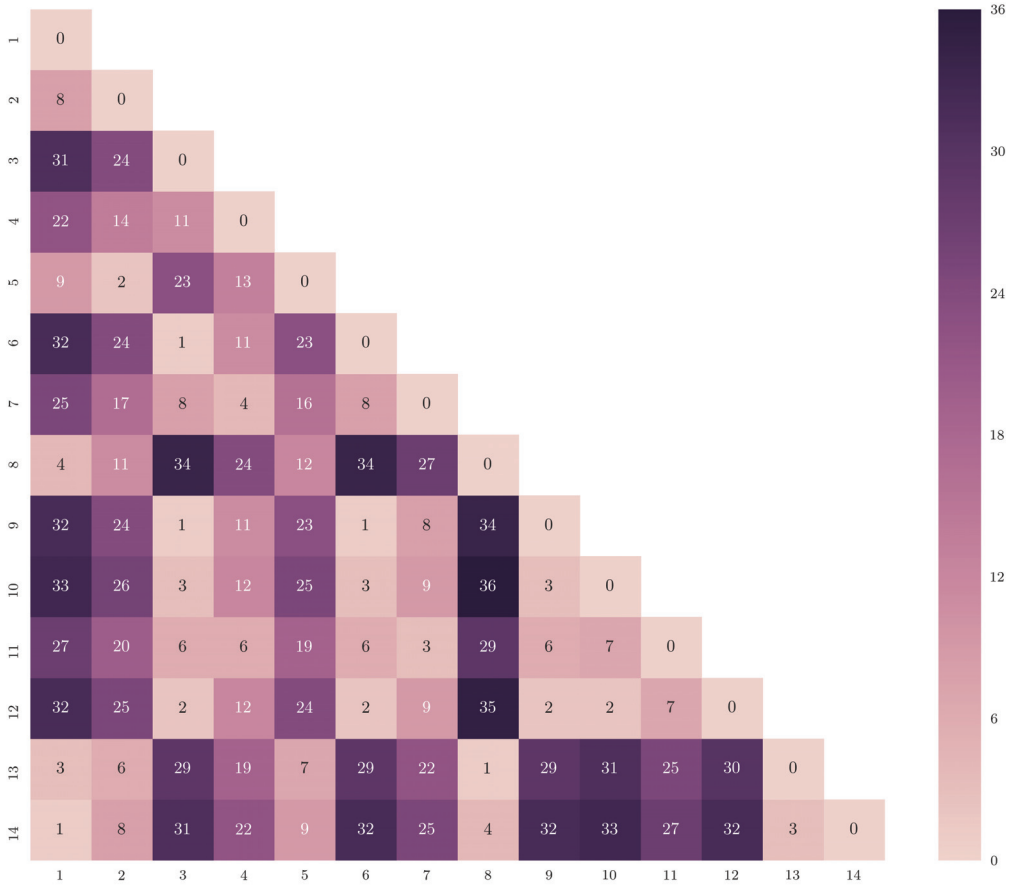


Figure 24: Similarity matrix of each test based on the Euclidean distance from the centroid of each test cluster.

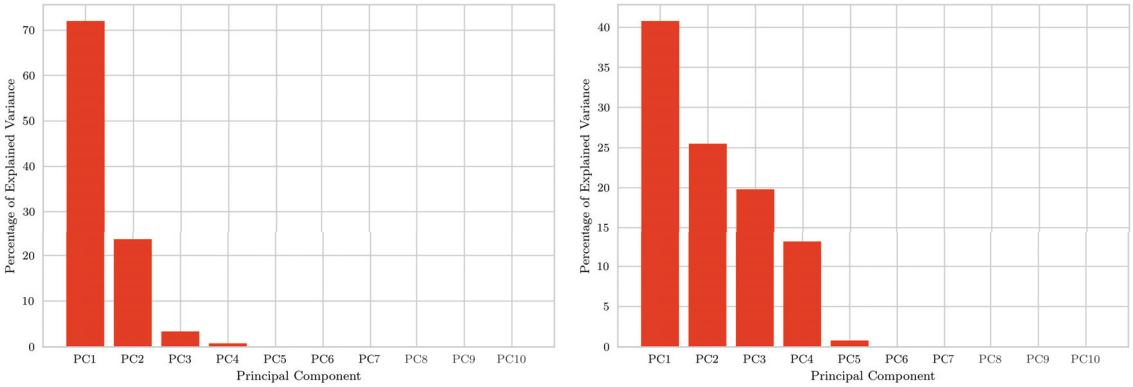
Table 3: Pairwise test comparison of RMSE of measured temperatures of a few selected components.

Test pair	Component 1	Component 2	Component 3
1 (Test 14 vs Test 1)	0.3 °C	1.4 °C	0.4 °C
2 (Test 13 vs Test 8)	0.8 °C	0.9 °C	0.4 °C
3 (Test 12 vs Test 6)	2.0 °C	0.1 °C	0.7 °C
4 (Test 11 vs Test 7)	5.8 °C	0.1 °C	1.3 °C

4.2 Outlier detection

This test was performed with the validation data, or the observations of tests 14, 13, 12 and 11. First, the training data were inputted into the model to establish a baseline of normal data for the error ellipse. Then the validation data were inputted. The observations include data from faulty sensors in two cases: temperature sensors failing in the middle of the test and failed temperature sensors at the beginning of the tests. In the first case, no signals were transmitted, while in the other case, the signals from the sensor are noise.

Conducting PCA on the time series data captures the behaviour of measurements originating from a faulty sensor compared to faultless data. With the existence of a clear outlier in the data, the spread of variance is more significant. The impact of the outlier is observable from the spread of variance in Figures 25a and 25b. With the presence of an outlier in the data, the variance is spread across multiple PC vectors, unlike in Figure 22 where the first PC vector captured 99% of the variance. In the case of a faulty sensor that transmits no signals, 85% of the data can be explained with the first two PC vectors, as seen in Figure 25a. However, with the presence of multiple failed sensors, the variance is spread across multiple PC vectors, as seen in Figure 25b. In the case of sensors omitting noise, three PC vectors are needed to explain 85% of the data, and the noisy signals increase the variance significantly, spreading it across multiple PC vectors.



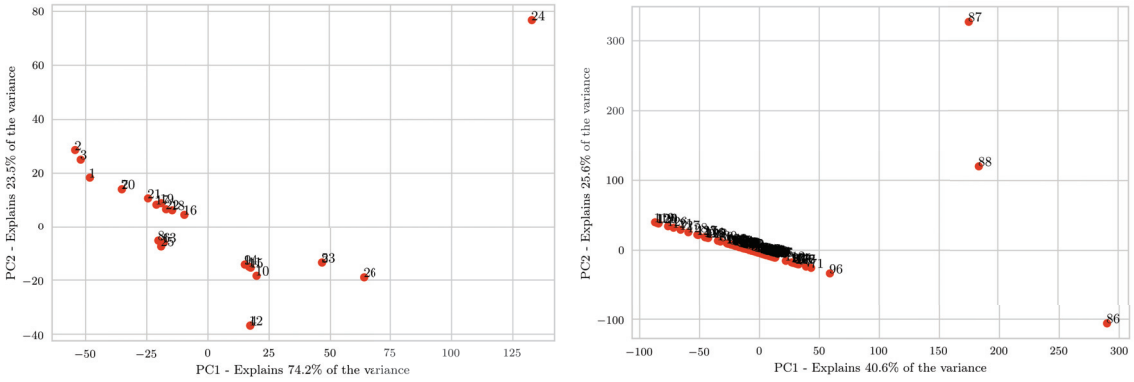
(a) one failed sensor.

(b) multiple failed sensors.

Figure 25: Explained variance ratio per each principal component in different failure cases from different dataset.

The observations are projected into two-dimensional space in Figures 26a and 26b. The mean and variance of the observations originating from a faulty sensor deviate from the rest of the observations. When not considering the outliers, the rest of the observations are contained in an ellipse. The faulty sensor in Figure 26a resembles data measured from component number 24. In Figure 26b, the faulty sensors were measuring components numbered 86, 87 and 88. Performance of the

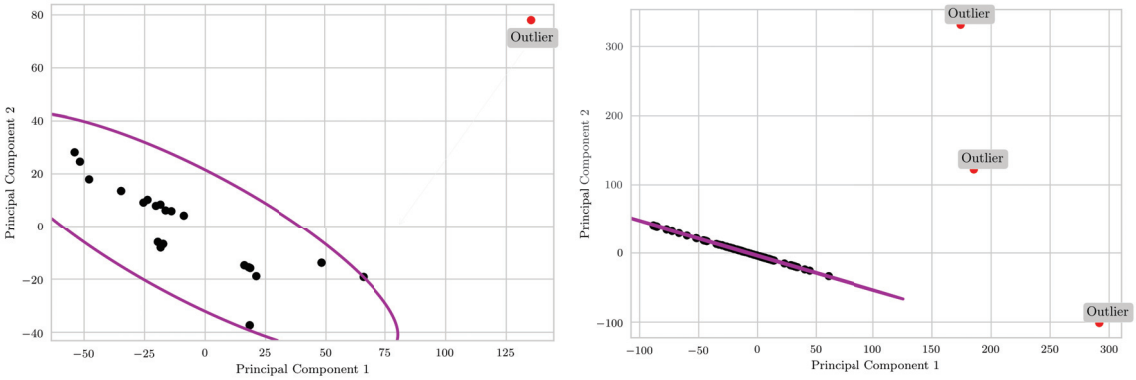
outlier detection algorithm is observable in Figures 27a and 27b. The algorithm detects accurately all the observations originating from the faulty sensors and the data originating from faultless sensors are captured inside an ellipse.



(a) one failed sensor.

(b) multiple failed sensors.

Figure 26: Data containing outliers from two failure cases projected into 2D space.



(a) one failed sensor.

(b) multiple failed sensors.

Figure 27: Error ellipse in an outlier detection in two failure scenarios.

4.3 Temperature estimation

The results of MC-based temperature estimation for the validation data is showcased. The different system states and transition probabilities are obtained and computed from historical time series data to determine the MC matrix. Employing the MC to model the system development is depicted in Figure 28. The states and transition probabilities are obtained from historical data, and the system development is captured by simulating a VSTF. For instance, consider the system observed at state 71.5 °C. When the system is observed again, the next state is either 71.5 °C or 71.6 °C. Transitioning to either of these states is governed by their probabilities. If a transition to state 71.5 °C has occurred, the next possible transitions are 71.5 °C, 71.6 °C or 71.7 °C. Time series are then simulated by sampling N number of VSTF. In the simulation, the states are calculated to one decimal place.

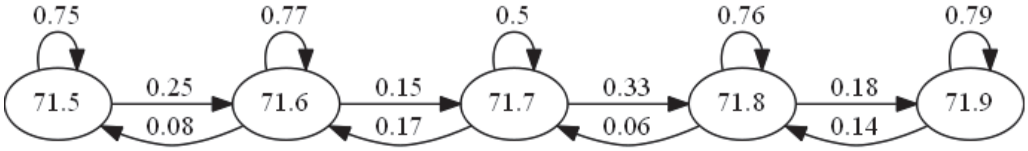


Figure 28: Sample of historical transition probabilities of a selected component from state 71.5 °C to 71.9 °C.

First, a comparison between different simulation sample-sizes is given. The numbers of simulated time series for one component are 10, 100, 1000 and 10000, and the results are shown in Table 4. The mean error and 95% confidence interval are displayed, and no significant accuracy has been achieved with larger sample-sizes. The difference between 10 and 10000 sampling is 30 minutes of computation time. MC sampling is computationally heavy, and a larger sample increases accuracy. However, $N = 100$ sampling is the optimal choice between accuracy and computation time. VSTF with MC is illustrated in Figure 29 and a histogram obtained by MC sampling is shown in Figure 30.

Table 4: Test statistics of temperature estimation for one component with different sample-sizes.

N	$5\%ile$	$95\%ile$	M	MAD
10	71.4	74.7	73.9	1.6
100	71.9	74.6	73.8	0.9
1000	72.0	74.6	73.6	0.9
10000	72.0	74.6	73.6	0.8

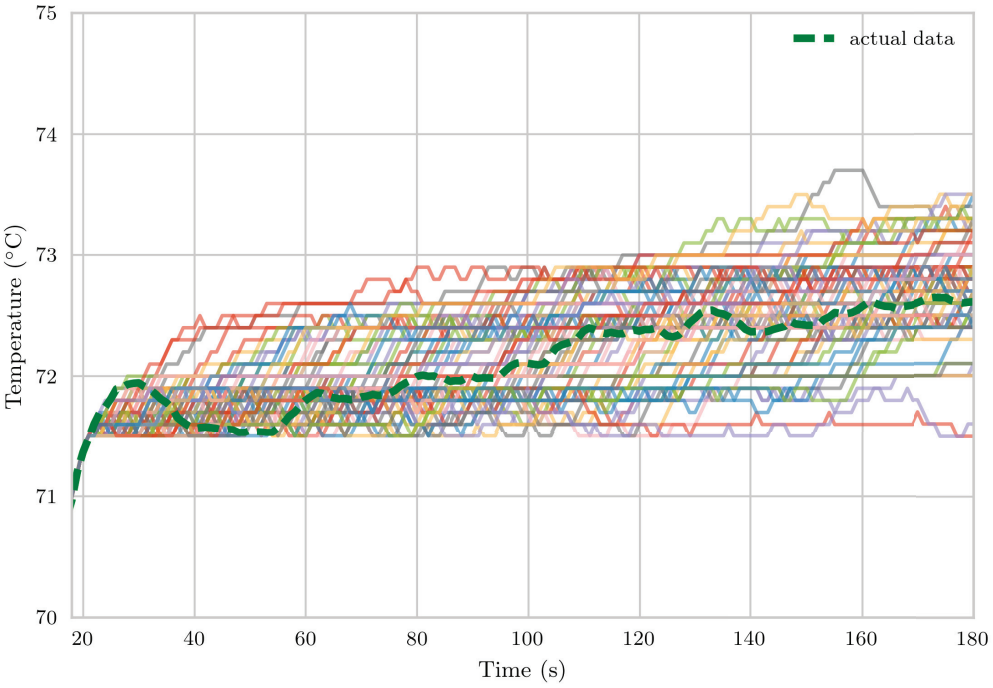


Figure 29: 100 simulated time series from a stationary distribution of one component.

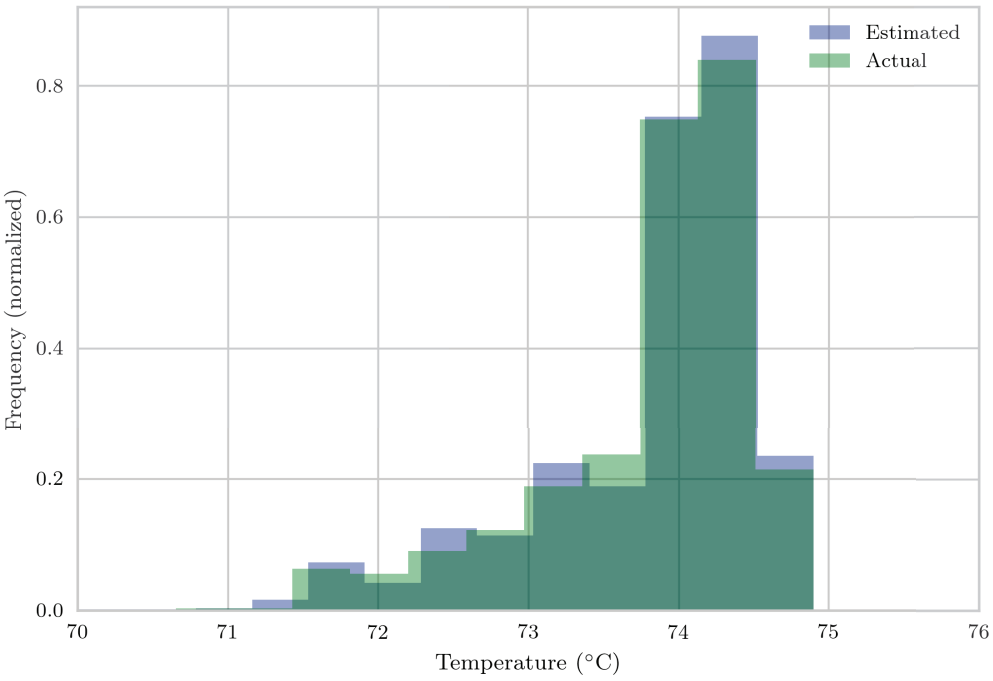


Figure 30: Distribution derived from 100 simulated time series.

4.3.1 Case 1: Sensor failure before stabilization

In this case, it is of utmost importance to estimate stability time and stability temperature. The estimates were obtained by first calculating the stationary distribution of MC and later simulating 100 time series and studying their distributions. The simulated estimates were then compared to the historical data of stability time and stability temperature in Table 5. Only the results of the best matching pair and the worst matching pair are presented. The results are consistent with the earlier observation of the similarity between the test, as shown in Table 3.

Table 5: Simulation results for Test pair 1.

	Component 1			Component 2			Component 3		
	Act	Est	RMSE	Act	Est	RMSE	Act	Est	RMSE
t_{stab}	75 min	71 min	6%	60 min	56 min	5%	90 min	87 min	3%
T_{stab}	79° C	76° C	3%	82° C	79° C	4%	92° C	91° C	1%

Table 6: Simulation results for Test pair 4.

	Component 1			Component 2			Component 3		
	Act	Est	RMSE	Act	Est	RMSE	Act	Est	RMSE
t_{stab}	75 min	58 min	22%	57 min	44 min	22%	82 min	61 min	25%
T_{stab}	75° C	72° C	4%	82° C	76° C	7%	85° C	80° C	6%

4.3.2 Case 2: Sensor failure after stabilization

This test case analyses how quickly the model can foresee a trend and exceed a possible threshold. Based on the example in Figure 31, the actual empirical temperature is plotted with data until time t and after time t with a lighter dotted line. The threshold limit is coloured with a vertical line, and the temperature surpasses the threshold at time $t + m$. The next section considers how quickly the model can foresee the surpassing of the threshold.

VSTF was sampled every 10 s after the sensor fault occurred until time $t+n$ when the possibility of reaching the threshold given the current temperature estimate was greater than 5%. Preferably, the time $t + n$ should be below $t + m$. The probability for continuing the inverter operating were set as:

$$\mathrm{P}(\text{Exceeding threshold} \mid \text{Current temperature}) \leq 5\% \,. \tag{36}$$

The choice of 5% probability was arbitrary, but the use of 5%ile and 95%ile is widely agreed upon in the literature [51]. If the probability is larger than 5%, then the run is aborted.

Figure 31 shows an example of simulated system failure. Projection of the time series data after the failure at $t = 150$ s is presented as a reference. The failure threshold is set to an arbitrary value of 73 °C, and according to the projection the threshold is exceeded at $t = 260$ s. Different system developments were simulated, and the probability of exceeding the threshold was evaluated.

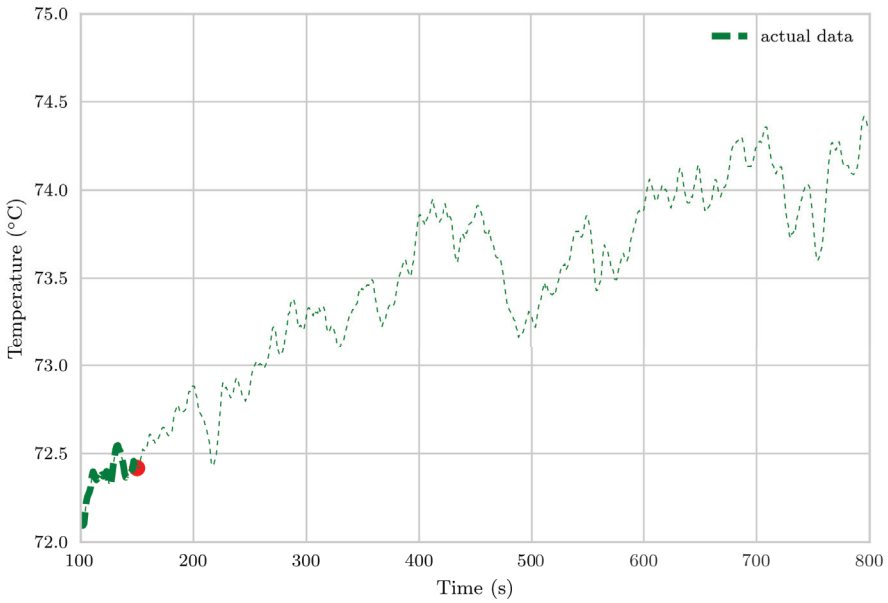


Figure 31: Example of simulated sensor failure at time $t = 150$ s and projected data as reference.

Figure 32 presents an example of the time instant when the model anticipated exceeding the 73 °C threshold in comparison to the actual threshold exceeding time of $t = 260$ s. Three component pairs were evaluated, the best matching pair and two of the worst matching pairs according to the test statistics. Component 3 from test pair 1 represents the best fit. Component 3 from test pair 3 and component 1 from test pair 4 are considered the two worst scenarios, representing significant under-fitting and over-fitting, respectively. Figure 33 represents the two cases of over- and under-fitting, while Figure 29 illustrates a good fit. In theory, over-fitting results in higher temperatures, which yields to anticipating exceeding the threshold earlier. Under-fitting, in contrast, gives lower temperature estimates than the actual values. Therefore, the threshold could be exceeded before the model anticipates it.

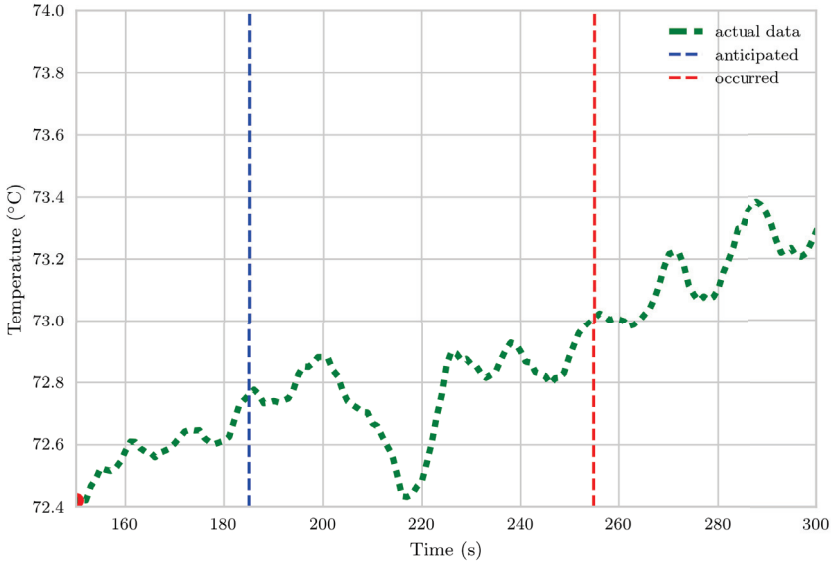
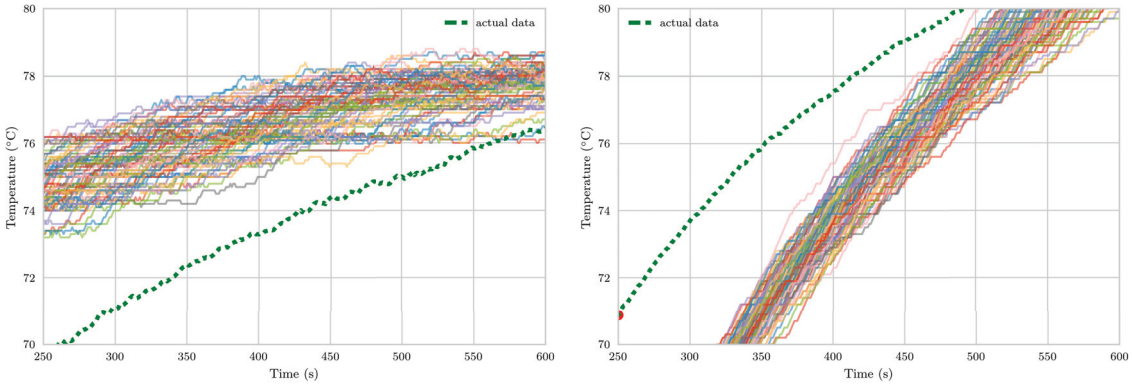


Figure 32: Time of forecasted threshold exceeding.



(a) Over-fitted estimates.

(b) Under-fitted estimates.

Figure 33: Over-fitted estimates and under-fitted estimates.

Only VSTF horizon was considered and the results are presented in Table 7. In all cases, exceeding the threshold was anticipated beforehand, illustrating the robustness of the proposed method. A failed case in which the anticipation time is after the threshold exceeding time could not be simulated with the historical data. In three cases, the model anticipated exceeding the threshold before the time occurrence. Case 1 refers to component number 3 from test pair number 1, case 2 is component number 1 from test pair 4 and case 3 is component number 3 from test pair 3.

Table 7: Time of occurred threshold exceeding vs. the time threshold exceeding was anticipated. The anticipated time is a result of 100 VST simulations. Time is counted from the moment of sensor failure.

Case	Occurred	Anticipated	SD
1	105 s	35 s	1 s
2	314 s	13 s	2 s
3	79 s	21 s	1 s

5 Conclusions

The purpose of this work was to study the possibility of identifying a broken temperature sensor in an outdoor inverter and estimating the temperature of a single component and evaluating if the inverter can continue operating when the sensor is broken. The historical data for this study are temperature time series data measured during different experiments of a 2-MW outdoor central inverter prototype conducted in a laboratory. Information about the internal system parameters of the inverter was not available.

Statistical evaluation of the data ruled out curve fitting and time series analysis as means for temperature estimation. Moreover, there were uncertainties in temperature data due to inherent and interference noise. Temperature distribution, stabilization temperature and time were also different for each component. However, the statistical evaluation showed that there are similarities between different experiments, suggesting fairly similar operating conditions. The difference of temperature distribution in each component and similarities between experiments made a data-driven approach the only feasible option for temperature estimation. The large size of the data matrices indicated that MVA should be applied in the proposed model.

A new approach for detecting a faulty temperature sensor and temperature estimation of PV inverter was presented. The approach utilizes only temperature time series data and consists of an outlier detection block and an estimation block. The outlier detection block analyses time series data, computes and saves data statistics to DB and detects faulty temperature sensors. PCA were used to transform the input data to a low-dimensional space. K-means algorithm was chosen for its simplicity to analyse the clustering of the observations. Observations were clustered and their similarities ranked by Euclidean distance to later establish a priori knowledge for the estimation block. A density ellipse with a cost function is drawn in the lower-dimension space based on empirical covariance from the observations to detect outliers residing outside the ellipse. The estimation block was based on Markov chain (MC) since there was strong theoretical basis for using it to model the system, and it has been used successfully in many PV system power output forecasting applications. Then, a priori data returned from outlier detection block were used to establish a transition matrix and later computed the stationary distribution. Also, 100 temperature time series were simulated from the MC stationary distribution. The distributions were then analysed to estimate the t_{stab}, T_{stab} and the probability of exceeding a predetermined temperature threshold.

First, large sets of time series data without outliers were inputted into the model to establish a baseline and build a DB. Next, four different data-sets with outliers were inputted into the model separately for validation. Outlier detection, temperature time series estimates and threshold exceeding anticipation were validated. All outliers were detected by the proposed methodology, and each dataset was assigned a priori data from the DB. Three components were selected to evaluate pairwise

RMSE of temperature time series between the four data-sets and their established a priori. The RMSE of the estimation block for the best performing combination of test pairs for t_{stab} and T_{stab} parameters were -3% and -1%, respectively, while the worst matching pair had RMSE values of -25% and -6% for t_{stab} and T_{stab}. In all cases, the model anticipated exceeding the threshold ahead of time, even in cases of over- and under-fitting.

The goals of the thesis were achieved. As was suggested in the literature, the proposed method had to be built on a combination of several different approaches. The proposed method required no knowledge of the PV inverter. Based only on the input temperature time series data, the model managed to categorize the components into correct clusters representing different operating points. Failed sensors were always identified from the time series data. The results reveal that in the absence of a faulty sensor, the first and second PC vector can cover 99% of the variance. Including outliers, the variance spreads significantly to other PC vectors. The error ellipse method proved to be simple and effective. Similar error ellipse methods were not found in the reviewed literature; thus, the method is new and unique. Dimensionality reduction with PCA was essential for the clustering algorithm to work. Moreover, the proposed outlier detection method is graphical and based on the empirical covariance and projecting the data into the lower dimension is computationally efficient.

The VST forecast results were consistent with low standard deviation as verified by executions of 100 iterations. The advantage of the developed model is that it can be used for estimating the temperature of the component based only on temperature time series data. Moreover, the model provides probabilities for risk management for operating the inverter. Even when the estimates were clearly over- and under-fitted compared to the actual data, the model anticipated the threshold exceeding beforehand. The results of the proposed MC-based method suggest that the method can be used for damage prediction and risk analysis when operating the inverter in the absence of a temperature sensor. This study introduced the use of robust statistics in the estimation procedure, which has not been done in the previous studies. Robust statistics alleviated the uncertainties of the simulations in achieving good t_{stab} and T_{stab} estimates.

The model is adaptive and simple. PCA was shown in reviewed studies to be important, and it proved to be an essential part of the method. Using MC was simple and effective. The proposed MC method is similar to Markov Chain Monte Carlo (MCMC) methods seen in various studies in that they both generate a simulated distribution from samples. Since the available data were limited, random sampling to generate simulated data was crucial for estimating system behaviour and evaluating uncertainty propagation. MCMC simulation methods in the reviewed studies involved sampling 10000 samples, which is computationally heavy [51]. In this study, the simulation results showed that sampling 100 samples yielded the same accuracy as 10000 samples and required significantly less computational

time. MC is not dependent on internal system knowledge, nor does it make distribution assumptions. However, the proposed model does require a priori knowledge of the component temperature data. As the results show, with a good a priori, the temperature time series estimates are accurate, and the error statistics are low. The K-means algorithm was sufficient for determining a reasonable a priori. With more historical data, the temperature time series gets more precise. As emphasized in the reviewed studies, successful MC implementation depends on the availability of historical data over long observation periods and on accurate classification of the historical data of the PV system according to their operating points. Compared to output forecasting results in the reviewed studies, the derived temperature estimates and the error statistics in the case of the best matching pair were low. The results are promising, and the proposed model can be developed further for real-time PV inverter application.

Although the proposed outlier detection method detected all the outliers when tested under real-time data, the accumulation of real-time data makes the outlier detection slower. Since the method is based on robust statistics, it took time for the outlier to deviate outside the error ellipse. In the worst instance, the algorithm took 2.5 minutes to detect a sensor failure from the accumulation of 3 hours' worth of data. The proposed error ellipse method is not optimized for real-time performance, and the error ellipse method is limited by the dimensionality of the lowered space. The method is geometrical, and it works only in two and three dimensions. Furthermore, the method does not work if the number of PC vectors that cover 85% of the data is greater than three.

The method assumes constant operating conditions. The proposed model did not consider variation in the ambient temperature. The projected estimates, therefore, are applicable only to operating conditions that share the same ambient temperature at which the raw data were collected. Nevertheless, raw data could be collected under different weather conditions, and a categorically different MC could be built. Furthermore, given the properties of MC, the forecast model should be applied to VST predictions. MC modelling is not reasonable for long-term yearly predictions, because over the long term, there can be noticeable seasonal trends, and the memoryless property might not be fulfilled. Moreover, it was assumed that the components remained faultless throughout the operation of the inverter and a faulty temperature sensor did not indicate a faulty component. Additionally, the cooling system was assumed to be unobstructed and operating flawlessly. In reality, the cooling system can be obstructed with an accumulation of dust, making the airflow suboptimal and thereby raising the temperatures. The applied approach of eigenvalue decomposition to calculate PCA and DTMC stationary distribution was not optimized for low memory consumption. Since the proposed model is data-driven, it relies heavily on the quantity and quality of historical data.

The outlier detecting block is good for detecting anomalies. For example, during IDA, temperature deviations of identical components were deemed outliers due to

cooling system obstructions and poor current dividing schemes. This result suggests that the outlier detection methodology can be applied within the design phase of the inverter to easily detect small deviations missed by the engineers and to quantify the performance of different components and topologies. The methodology can be integrated into the inverter software to continuously monitor the fitness of a component to improve inverter reliability. Deviations from the normal conditions can be used in a pre-emptive manner to schedule maintenance before deterioration or failure occurs.

Considering that the temperature of the components is directly affected by the ambient temperature, more efforts are needed to study the effect of varying ambient temperature and integrate it into the proposed model. This can be achieved by incorporating more variables and historical data into the model since the methodology is data-driven. Data should be recorded from field operation as well as operating the inverter under different operating point combinations. Moreover, the measurements should be recorded before the moment of power conversion until power conversion is finished, and not mid-operation after the power conversion has already started.

With the accumulation of more field data and incorporating more variables into the model, an automated approach should be studied. As the number of variables and observations grows, it is not possible to analyse the relationship between all the variables across all the observations in an insightful way by a person. Clearly, an automated approach is needed for outlier detection and mapping the relationship between different operating point variables, ambient temperature, the cooling system and to the different component temperatures. Another important topic for future research is to further improve the computational efficiency of the proposed models for real-time PV inverter application.

References

[1] N. Mohan and T. Undeland. *Power electronics: converters, applications, and design*. Wiley, 2007. ISBN: 9788126510900.

[2] E. Perez et al. "Predictive power control for PV plants with energy storage". In: *IEEE Transactions on Sustainable Energy* 4.2 (2013), pp. 482–490.

[3] M. Rekinger et al. "Connecting the sun: solar photovoltaics on the road to large-scale grid integration". In: *European Photovoltaic Industry Association, Brussels, Belgium, Full report* (2012).

[4] S. Kouro et al. "Grid-connected photovoltaic systems: An overview of recent research and emerging PV converter technology". In: *IEEE Industrial Electronics Magazine* 9.1 (2015), pp. 47–61.

[5] J. Antonanzas et al. "Review of photovoltaic power forecasting". In: *Solar Energy* 136 (2016), pp. 78–111.

[6] R. J. Wai and W. H. Wang. "Grid-connected photovoltaic generation system". In: *IEEE Transactions on Circuits and Systems* 55.3 (2008), pp. 953–964.

[7] A. K. Barnes, J. C. Balda, and A. Escobar-Mejıa. "A semi-Markov model for control of energy storage in utility grids and microgrids with PV generation". In: *IEEE Transactions on Sustainable Energy* 6.2 (2015), pp. 546–556.

[8] C. Lupangu and R. Bansal. "A review of technical issues on the development of solar photovoltaic systems". In: *Renewable and Sustainable Energy Reviews* 73 (2017), pp. 950–965.

[9] R. Teodorescu, M. Liserre, and P. Rodriguez. *Grid Converters for Photovoltaic and Wind Power Systems*. Wiley, 2011. ISBN: 9781119957201.

[10] S. Shimura et al. "Production costs estimation in photovoltaic power plants using reliability". In: *Solar Energy* 133 (2016), pp. 294–304.

[11] P. Hacke et al. "A status review of photovoltaic power conversion equipment reliability, safety, and quality assurance protocols". In: *Renewable and Sustainable Energy Reviews* 82 (2018), pp. 1097–1112.

[12] Y. Song and B. Wang. "Survey on reliability of power electronic systems". In: *IEEE Transactions on Power Electronics* 28.1 (2013), pp. 591–604.

[13] S. Yang et al. "An industry-based survey of reliability in power electronic converters". In: *IEEE transactions on Industry Applications* 47.3 (2011), pp. 1441–1451.

[14] H. Wang, M. Liserre, and F. Blaabjerg. "Toward reliable power electronics: Challenges, design tools, and opportunities". In: *IEEE Industrial Electronics Magazine* 7.2 (2013), pp. 17–26.

[15] J. Falck, M. Andresen, and M. Liserre. "Active methods to improve reliability in power electronics". In: *IECON 2017 - 43rd Annual Conference of the IEEE Industrial Electronics Society*. IEEE. 2017, pp. 7923–7928.

[16] J. Shackelford. *Introduction to Materials Science for Engineers, Global Edition*. Pearson Education Limited, 2015. ISBN: 9780273793984.

[17] C. Chatfield. *Time-series forecasting*. CRC Press, 2000. ISBN: 9781584880639.

[18] ABB. *ABB central inverters PVS980 – 1818 to 2091 kVA*. Brochure. 2018. URL: https://library.e.abb.com/public/bbbc00f6b0ad4d3f9a703a9a049d53e9/PVS980_central_inverters_flyer_3AXD50000027473_RevJ_EN_lowres.pdf.

[19] ABB. *Now available: ABB's PVS980, best in class central inverter*. Brochure. 2016. URL: http://www04.abb.com/global/seitp/seitp202.nsf/0/84b9e857cadd59acc1257fd40028aafb/%5C$file/EN_Solar+inverter_PVS980.pdf.

[20] J. Fraden. *Handbook of Modern Sensors: Physics, Designs, and Applications*. Springer, 2015. ISBN: 9783319193038.

[21] BIPM et al. *IUPAP, and OIML, 2008,"Evaluation of Measurement Data—Guide to the Expression of Uncertainty in Measurement," Joint Committee for Guides in Metrology*. Tech. rep. Technical Report No. JCGM 100, 2008.

[22] T. A. N. Heirung et al. "Stochastic model predictive control—how does it work?" In: *Computers & Chemical Engineering* (2017).

[23] R. Kadri, J. P. Gaubert, and G. Champenois. "An improved maximum power point tracking for photovoltaic grid-connected inverter based on voltage-oriented control". In: *IEEE transactions on industrial electronics* 58.1 (2011), pp. 66–75.

[24] L. Ying Zi, N. Jin Cang, L. Ru, et al. "Optimal control for dynamic grid connected photovoltaic system based on Markov chain". In: *2007 International Conference on Electrical Machines and Systems*. IEEE. 2007, pp. 227–231.

[25] R. Sayed, Y. Hegazy, and M. Mostafa. "Modeling of photovoltaic based power stations for reliability studies using Markov chains". In: *2013 International Conference on Renewable Energy Research and Applications*. IEEE. 2013, pp. 667–673.

[26] Y. Z. Li and J. C. Niu. "Forecast of power generation for grid-connected photovoltaic system based on Markov chain". In: *Power and Energy Engineering Conference, 2009. APPEEC 2009. Asia-Pacific*. IEEE. 2009, pp. 1–4.

[27] S. Särkkä. *Bayesian filtering and smoothing*. Vol. 3. Cambridge University Press, 2013. ISBN: 9781107619289.

[28] G. James et al. *An introduction to statistical learning*. Vol. 112. Springer, 2013. ISBN: 9781461471370.

[29] M. Kuhn and K. Johnson. *Applied predictive modeling*. Vol. 26. Springer, 2013. ISBN: 9781461468486.

[30] S. Khomfoi, L. M. Tolbertt, and B. Ozpineci. "Cascaded H-bridge multilevel inverter drives operating under faulty condition with AI-based fault diagnosis and reconfiguration". In: *2007 IEEE International Electric Machines Drives Conference*. Vol. 2. IEEE. 2007, pp. 1649–1656.

[31] A. A. Silva, A. M. Bazzi, and S. Gupta. "Fault diagnosis in electric drives using machine learning approaches". In: *2013 International Electric Machines Drives Conference*. IEEE. 2013, pp. 722–726.

[32] T. Wang et al. "Cascaded H-bridge multilevel inverter system fault diagnosis using a PCA and multiclass relevance vector machine approach". In: *IEEE Transactions on Power Electronics* 30.12 (2015), pp. 7006–7018.

[33] W. F. Godoy et al. "An application of artificial neural networks and PCA for stator fault diagnosis in inverter-fed induction motors". In: *2016 XXII International Conference on Electrical Machines*. IEEE. 2016, pp. 2165–2171.

[34] J. Martins et al. "Fault detection and diagnosis of grid-connected power inverters using PCA and current mean value". In: *IECON 2012 - 38th Annual Conference on IEEE Industrial Electronics Society*. 2012.

[35] L. Wasserman. *All of Statistics: A Concise Course in Statistical Inference*. Springer, 2004. ISBN: 9780387402727.

[36] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, 2013. ISBN: 9780387216065.

[37] H. T. Pedro and C. F. Coimbra. "Assessment of forecasting techniques for solar power production with no exogenous inputs". In: *Solar Energy* 86.7 (2012), pp. 2017–2028.

[38] F. Chan and H. Calleja. "Reliability estimation of three single-phase topologies in grid-connected PV systems". In: *IEEE Transactions on Industrial Electronics* 58.7 (2011), pp. 2683–2689.

[39] S. Yang et al. "Condition monitoring for device reliability in power electronic converters: A review". In: *IEEE Transactions on Power Electronics* 25.11 (2010), pp. 2734–2752.

[40] M. Q. Raza, M. Nadarajah, and C. Ekanayake. "On recent advances in PV output power forecast". In: *Solar Energy* 136 (2016), pp. 125–144.

[41] S. Sobri, S. Koohi-Kamali, and N. A. Rahim. "Solar photovoltaic generation forecasting methods: A review". In: *Energy Conversion and Management* 156 (2018), pp. 459–497.

[42] A. Ragnacci et al. "Exploiting dimensionality reduction techniques for photovoltaic power forecasting". In: *2012 IEEE International Energy Conference and Exhibition*. IEEE. 2012, pp. 867–872.

[43] S. Qijun et al. "Photovoltaic power prediction based on principal component analysis and Support Vector Machine". In: *2016 IEEE Innovative Smart Grid Technologies-Asia*. IEEE. 2016, pp. 815–820.

[44] M. Malvoni, M. G. De Giorgi, and P. M. Congedo. "Photovoltaic forecast based on hybrid PCA–LSSVM using dimensionality reducted data". In: *Neurocomputing* 211 (2016), pp. 72–83.

[45] S. S. Soman et al. "A review of wind power and wind speed forecasting methods with different time horizons". In: *North American power symposium, 2010*. IEEE. 2010, pp. 1–8.

[46] R. H. Inman, H. T. Pedro, and C. F. Coimbra. "Solar forecasting methods for renewable energy integration". In: *Progress in energy and combustion science* 39.6 (2013), pp. 535–576.

[47] A. Sharma and A. Kakkar. "Forecasting daily global solar irradiance generation using machine learning". In: *Renewable and Sustainable Energy Reviews* (2017).

[48] T. Hong and S. Fan. "Probabilistic electric load forecasting: A tutorial review". In: *International Journal of Forecasting* 32.3 (2016), pp. 914–938.

[49] G. Graditi, S. Ferlito, and G. Adinolfi. "Comparison of Photovoltaic plant power production prediction methods using a large measured dataset". In: *Renewable Energy* 90 (2016), pp. 513–519.

[50] S. K. Chow, E. W. Lee, and D. H. Li. "Short-term prediction of photovoltaic energy generation by intelligent approach". In: *Energy and Buildings* 55 (2012), pp. 660–667.

[51] S. Talari et al. "Stochastic modelling of renewable energy sources from operators' point-of-view: A survey". In: *Renewable and Sustainable Energy Reviews* (2017).

[52] P. Zhang et al. "Reliability assessment of photovoltaic power systems: Review of current status and future perspectives". In: *Applied Energy* 104 (2013), pp. 822–833.

[53] V. Kulkarni. *Introduction to Modeling and Analysis of Stochastic Systems*. Springer, 2010. ISBN: 9781441917720.

[54] R. Durrett and R. Durrett. *Essentials of stochastic processes*. Springer, 2016. ISBN: 9781461436157.

[55] M. J. Sanjari and H. Gooi. "Probabilistic Forecast of PV Power Generation Based on Higher Order Markov Chain". In: *IEEE Transactions on Power Systems* 32.4 (2017), pp. 2942–2952.

[56] Y. Li, L. Ru, and J. Niu. "Power structure optimization for grid-connected photovoltaic system based on Markov decision processes". In: *2011 6th IEEE Conference on Industrial Electronics and Applications*. IEEE. 2011, pp. 2688–2692.

[57] J. F. Junior et al. "Forecasting regional photovoltaic power generation-a comparison of strategies to obtain one-day-ahead data". In: *Energy Procedia* 57 (2014), pp. 1337–1345.

[58] J. Xiu, C. Zhu, and Z. Yang. "Prediction of solar power generation based on the principal components analysis and the BP neural network". In: *2014 IEEE 3rd International Conference on Cloud Computing and Intelligence Systems*. IEEE. 2014, pp. 366–369.

[59] A. Mellit and S. A. Kalogirou. "Artificial intelligence techniques for photovoltaic applications: A review". In: *Progress in energy and combustion science* 34.5 (2008), pp. 574–632.

[60] A. Field. *Discovering statistics using IBM SPSS statistics*. Sage, 2013.

[61] R. Shumway and D. Stoffer. *Time Series Analysis and Its Applications: With R Examples*. Springer New York, 2010. ISBN: 9781441978646.

[62] G. Shrestha and L. Goel. "A study on optimal sizing of stand-alone photovoltaic stations". In: *IEEE Transactions on Energy Conversion* 13.4 (1998), pp. 373–378.

[63] G. Zini, C. Mangeant, and J. Merten. "Reliability of large-scale grid-connected photovoltaic systems". In: *Renewable Energy* 36.9 (2011), pp. 2334–2340.

[64] H. M. Taylor and S. Karlin. *An introduction to stochastic modeling*. Academic press, 2014. ISBN: 9780233814162.

[65] M. DeGroot and M. Schervish. *Probability and Statistics*. Addison-Wesley, 2002. ISBN: 9780201524888.

[66] C. J. Geyer. "Breakdown point theory notes". In: *Class Notes on Nonparametric Statistics* (2006).

[67] K. Mardia, J. Kent, and J. Bibby. *Multivariate analysis*. Academic Press, 1979. ISBN: 9780124712508.

[68] A. C. Rencher. *Methods of multivariate analysis*. Vol. 492. John Wiley & Sons, 2003. ISBN: 9780470178966.

[69] J. Marden. *Multivariate Statistics: Old School*. CreateSpace Independent Publishing Platform, 2015. ISBN: 9781456538835.

[70] M. Laatikainen. *Solar power generation forecast updated every hour*. URL: https://data.fingrid.fi/dataset/solar-power-generation-forecast-updated-every-hour.

[71] S. Seabold and J. Perktold. "Statsmodels: Econometric and statistical modeling with python". In: *9th Python in Science Conference*. 2010.

[72] J. Friedman, T. Hastie, and R. Tibshirani. *The elements of statistical learning*. Vol. 1. Springer, 2001. ISBN: 9780387216065.

[73] A. Kyprianou et al. "Definition and computation of the degradation rates of photovoltaic systems of different technologies with robust principal component analysis". In: *IEEE Journal of Photovoltaics* 5.6 (2015), pp. 1698–1705.

[74] G. Lawler. *Introduction to Stochastic Processes*. Taylor & Francis, 1995. ISBN: 9780412995118.

[75] J. Li et al. "Machine learning for solar irradiance forecasting of photovoltaic system". In: *Renewable Energy* 90 (2016), pp. 542–553.

[76] J. Kruschke. *Doing Bayesian data analysis: A tutorial with R.* Academic Press, 2014. ISBN: 9780124058880.

[77] Project Manager, ABB Product Group Solar. Private interview. May 2018.